# Nonparametric Triangular Simultaneous Equations Models with Weak Instruments[*]

Sukjin Han[†]

*Department of Economics*
*Yale University*

February 25, 2012

**Abstract**

Instrumental variables are widely used to identify and estimate causal relationships in models with endogenous explanatory variables. The consequences of weak instruments have been extensively studied in the literature with linear simultaneous equations models. One might conjecture that the problem of weak instruments becomes even more important when studying endogenous explanatory variables in nonparametric models, as more flexible models generally require stronger identification power, and hence plausibly stronger instruments.

This paper is the first to analyze the effect of weak instruments on identification, estimation, and inference in a nonparametric setting. We consider a triangular simultaneous equations model, and follow the control function approach for identification and estimation. We derive a necessary and sufficient rank condition for identification, based on which weak identification is established. Then, *nonparametric weak instruments* are defined as a sequence of reduced form functions that converges to a constant function. We characterize weak instruments as a multicollinearity problem or, more generally, as an inverse problem, which motivates the introduction of a regularization scheme. We propose a series estimation method with penalization to alleviate the effects of weak instruments. We derive the rate of convergence of the resulting penalized series estimator. Consistency and asymptotic normality are achieved with "mildly" weak instruments and a "rapidly" shrinking penalization parameter. Monte Carlo results show that the finite sample performance of the penalized estimator is appealing. The results of this paper are applied to an empirical example, where the effect of class size on test scores is estimated nonparametrically.

[*] I am very grateful to my advisors, Donald Andrews and Edward Vytlacil, and committee members, Xiaohong Chen and Yuichi Kitamura for their inspiration, guidance and support. I am deeply indebted to Donald Andrews for his thoughtful advice throughout the project. This paper has benefited from discussions with Joseph Altonji, Ivan Canay, Philip Haile, Keisuke Hirano, Han Hong, Joel Horowitz, Seokbae Simon Lee, Oliver Linton, Whitney Newey, Peter Phillips, Andres Santos, and Alex Torgovitsky. I gratefully acknowledge financial support from a Carl Arvid Anderson Prize of the Cowles Foundation.

[†] Email: sukjin.han@yale.edu. Tel: 1-203-623-6736. Comments are welcome.

*Keywords:* Triangular models, nonparametric identification, weak identification, weak instruments, series estimation, collinearity, penalization.
*JEL Classification Numbers:* C13, C14, C36.

# 1 Introduction

Instrumental variables (IVs) are widely used to identify and estimate models with endogenous explanatory variables. In linear simultaneous equations models, it is well-known that standard asymptotic approximations break down when instruments are weak in the sense that (partial) correlation between the instruments and endogenous variables is weak. The consequences of and solutions for weak instruments in linear settings have been extensively studied in the literature over the past decade; see, e.g., Bound, Jaeger and Baker (1995), Staiger and Stock (1997), Dufour (1997), Kleibergen (2002, 2005), Moreira (2003), Stock and Yogo (2005), and Andrews and Stock (2007).[1] Weak instruments in nonlinear parametric models have been studied less in the literature, either in the context of weak identification, e.g., by Stock and Wright (2000), Han and Phillips (2006), Newey and Windmeijer (2009), Andrews and Cheng (2010), or in a particular limited-dependent-variables version of simultaneous equations models by Han (2011).

One might expect that nonparametric models with endogenous explanatory variables will generally require stronger identification power than linear models as there is an infinite number of unknown parameters to identify, and hence stronger instruments may be required. Despite the problem's importance and the growing popularity of nonparametric models, weak instruments in nonparametric settings have not received much attention.[2] Also, surprisingly little attention has been paid to the consequences of weak instruments in applied research using nonparametric models. Part of the theoretical neglect is due to the existing difficulties embedded in nonparametric models. In a framework introduced in this paper, however, weak instruments can be formalized clearly and their effect can be analyzed without further difficulties.

This paper analyzes the effect of weak instruments on identification, estimation, and inference in a simple but widely-used nonparametric simultaneous equations model. We also provide estimation strategies that have desirable properties when instruments are possibly weak. Specifically, we consider a nonparametric triangular model. The model, which is fully described later, consists of a nonparametric structural equation (or outcome equation) $y = g(x) + \varepsilon$, where $g(\cdot)$ is a function of interest and $x$ is endogenous, and a nonparametric reduced-form equation $x = \Pi(z) + v$, where $z$ is a vector of instruments. This model is also considered in Newey, Powell and Vella (1999), where identification and estimation results are established in a situation without weak instruments.

[1] See Andrews and Stock (2007) for exhaustive survey of the literature on weak instruments.

[2] Chesher (2003) mentions the issue of weak instruments in applying his key identification condition in the empirical example of Angrist and Krueger (1991), and Blundell, Chen and Kristensen (2007) determine whether weak instruments are present in the Engel curve dataset of their empirical section. They do that by conducting the Stock-Yogo (2005) test developed in linear models applied to their reduced form which is linearized by sieve approximation.

We take the standard control function approach in the identification and estimation of the structural function $g(\cdot)$ as in Newey, Powell and Vella (1999): the model with conditional mean restrictions stated below implies that $E[y|x,z] = g(x) + E[\varepsilon|v] = g(x) + \lambda(v)$, where the endogeneity ($E[\varepsilon|v] \neq 0$) is controlled by introducing an additional unknown function $\lambda(v)$ of the reduced-form errors (or residuals in estimation). This equation serves as a basis for identification and estimation. The estimation method is series estimation because it is suitable for imposing the additive structure of the equation. Series estimation is also used in Newey, Powell and Vella (1999).

The nonparametric triangular model considered in this paper is frequently used in recent applied research such as Blundell and Duncan (1998) and Dustman and Meghir (2005), as it has a form analogous to its popular parametric counterpart. The series estimation based on the control function approach is also easy to implement in practice. More importantly, in analyzing weak instruments, the model has advantages over other nonparametric models with endogenous explanatory variables, such as the nonparametric IV (NPIV) model which is considered, e.g., in Newey and Powell (2003), Hall and Horowitz (2005) and Blundell, Chen and Kristensen (2007). The NPIV model is an alternative nonparametric model with different stochastic assumptions and no first-stage reduced-form equation. Unlike in the NPIV model, the specification of weak instruments is intuitive in the triangular model from the explicit reduced-form relationship. Also, clear interpretation of the effect of the weak instruments can be made from the series estimation of the implied equation derived above. Lastly, no other difficulties are intrinsic to the model, such as the ill-posed inverse problem which arises in the NPIV model.

The main contributions of the paper are summarized as follows. First, we derive novel identification results in nonparametric triangular models that complement existing results in the literature, and we establish the notion of weak identification based on these results. With a mild support condition, we show that a particular rank condition on $\Pi(\cdot)$ is necessary and sufficient for identification. This rank condition is substantially weaker than the sufficient rank condition established in Newey, Powell and Vella (1999). Deriving such a minimal rank condition is important in that a "slight violation" of it has a binding effect on identification and hence results in weak identification.

Second, the concept of *nonparametric weak instruments* is then defined, which generalizes the concept of weak instruments with a linear reduced form as in Staiger and Stock (1997). We consider sequences of reduced-form functions that converge to a *non-identification region*, namely, a space of reduced-form functions that violate the rank condition for identification. Under this localization, the signal diminishes relative to the noise in the system, and hence the model is weakly identified. In particular, we consider a sequence where the reduced-form functions become flatter. A particular rate is designated in terms of the sample size, which effectively measures the strength of the instruments and appears in our asymptotic results for the estimator of the structural function $g(\cdot)$.

In general, the weak instrument problem can be seen as an inverse problem. In the nonparametric control function framework, the problem becomes a nonparametric analogue of a multicollinearity problem. To see this, note that once the endogeneity is controlled by the control function, the model can be rewritten as an additive nonparametric regression $y = g(x)+\lambda(v)+\eta$ by defining $\eta = y-E[y|x, z]$. The endogenous variables $x$ and reduced-form errors $v$ comprise two regressors, but weak instruments result in the variation in $x$ being mainly driven from the variation of $v$, so that $x$ and $v$ are close to "collinear." Alternatively, a series representation of the regression equation reveals that the problem can be seen as where the regressors become less variable as the instruments become weak. This problem is related to the ill-posed inverse problem inherent to the NPIV model. The integral equation produced under the NPIV approach faces a discontinuity problem when recovering the structural function by inversion. Again, once the structural function is represented by a series approximation, this inverse problem is translated into a problem where the regressors have little variation, since they form the conditional expectation of basis functions (e.g., Kress (1989, p. 235)). The similarity of the problems motivates that regularization methods used in the NPIV literature to solve the ill-posed inverse problem can be introduced to our problem. There is, however, an important difference between the two problems in that, among the regularization methods used to solve the ill-posed inverse problem, only penalization alleviates the effect of weak instruments.

Third, given this insight, we introduce a penalization scheme in estimation as a regularization method to alleviate the effect of weak instruments. We define a penalized series estimator and establish its asymptotic properties. Our results on the rate of convergence of the estimator suggest the way in which weak instruments and penalization affect bias and variance. In particular, weak instruments characterized as a multicollinearity problem exacerbate bias and variance "symmetrically," unlike the situation in the linear regression model where multicollinearity results in imprecise estimates but does not introduce bias. Consistency and asymptotic normality are achieved provided that the instruments are only mildly weak and the penalization parameter shrinks sufficiently fast, which are discussed more rigorously later.

Penalization can reduce both bias and variance through the same mechanism working in an opposite direction to the effect of weak instruments. The finite sample performance of the penalized series estimator as measured by Monte Carlo simulations suggests that variance reduction is significant compared to an unpenalized estimator. Furthermore, despite the additional penalization term, the bias is no larger than that of the unpenalized estimators for a wide range of the strengths of instrument and magnitudes of penalization parameter, and in some cases even smaller.

This paper provides useful implications for applied researchers. First, when one is estimating a nonparametric structural function, the results of IV estimation and subsequent inference can be misleading even when the instruments are strong in terms of conventional criteria for linear models (such as the first-stage $F > 10$ in Staiger and Stock (1997)). Second, the theoretical result that weak instruments have a symmetric effect on bias and variance implies that the bias and variance trade-off is the same across different strengths of instruments, and hence weak instruments cannot be alleviated by the choice of the order of series. Third, penalization, on the other hand, can alleviate weak instruments by significantly reducing variance and sometimes bias. Fourth, the strength of instruments can be improved by having a nonparametric reduced form, so that the nonlinear relationship between the endogenous variable and instruments is fully exploited. Although a linear first-stage reduced form is commonly used in applied research,[3] it is more subject to weak instruments than is a nonparametric reduced form. Also, the nonparametric first stage estimation is not likely to worsen the overall convergence rate of the estimator, since the nonparametric rate from the second stage is already present. The issue of dimensionality of the instruments can be mitigated by using a single-index model or an additive model for the reduced-form. In the rare case where a linear reduced-form relationship is in fact justified by theory, one might need to be more cautious about weak instruments than in the case with a nonlinear reduced-form relationship.

We apply the findings of this paper to an empirical example, where we nonparametrically estimate the effect of class size on students' test scores. In a well-known paper by Angrist and Lavy (1999), the effect of class size on students' test scores is estimated using linear models. Class size is endogenous, and exogenous variation due to a rule on maximum class size is used as an instrument. In the present paper, we generalize the linear structural function of their model to be nonparametric. The flexible functional form allows the marginal effect to be different across class-size levels. Given this nonparametric extension, one can test whether the results of Angrist and Lavy (1999) are driven by parametric assumptions. We calculate penalized series estimates of the class-size effect, which indicates that the overall effect is negative while the marginal effect is diminishing. We also contrast (unpenalized) series estimates calculated based on linear reduced-form with those based on nonparametric reduced-form, and provide evidence that the instrument can be considered weak in a nonparametric sense. Lastly, with a larger sample that is also used in Horowitz (2011) where the IV is considered extremely strong, we compare our control function estimates with the estimates of Horowitz (2011) where the NPIV model is considered and hence the ill-posed inverse problem is present. Unlike under the NPIV approach, the estimate under the control function approach has a substantially narrow confidence band, which indicates that the data is informative about the class-size effect.

The rest of the paper is organized as follows. Section 2 introduces the triangular model and control function approach. Section 3 obtains new identification results for the model and discusses the resulting rank condition in the context of instrument relevance. Section 4 discusses lack of identification and weak identification, where the concept of nonparametric weak instruments is defined using the localization technique. Section 5 relates the weak instrument problem to the ill-posed inverse problem and motivates our penalized series estimator. The estimator is defined and the competing effects of weak instruments and penalization are discussed. Sections 6-7 establish the rate of convergence and consistency of the penalized series estimator and the asymptotic normality of some functionals of it. Section 8 presents Monte Carlo simulation results. Section 9 discusses the empirical application of estimating the effect of class size on test scores. Finally, Section 10 concludes.

[3] See, for example, Newey, Powell and Vella (1999), Blundell and Duncan (1998), Blundell, Duncan and Pendakur (1998), and Dustman and Meghir (2005).

# 2 Model

We consider a nonparametric triangular simultaneous equations model

$$\begin{aligned} y &= g_0(x, z_1) + \varepsilon, & \text{(2.1a)} \\ x &= \Pi_0(z) + v, & \text{(2.1b)} \\ E[\varepsilon|v, z] &= E[\varepsilon|v] \ \text{ a.s.}, & \text{(2.1c)} \\ E[v|z] &= 0 \ \text{ a.s.}, & \text{(2.1d)} \end{aligned}$$

where $g_0(\cdot, \cdot)$ is an unknown structural function of interest, $\Pi_0(\cdot)$ is an unknown reduced-form function, $x$ is a $d_x \times 1$ vector of endogenous variables, $z = (z_1, z_2)$ is a $(d_{z_1} + d_{z_2}) \times 1$ vector of exogenous variables and $z_2$ is a vector of excluded instruments. This model is also considered in Newey, Powell, and Vella (1999) (NPV). The stochastic assumptions (2.1c)-(2.1d) are often called the "control function" assumptions as they enable the control function approach to be employed, which is discussed below. The stochastic assumptions are more general than assuming full independence between $(\varepsilon, v)$ and $z$ and $E[v] = 0$. As NPV point out, the orthogonality condition (2.1c) allows for heteroskedasticity, whereas $(\varepsilon, v) \perp z$ does not. Without further conditions, assumptions (2.1c)-(2.1d) are not stronger nor weaker than $E[\varepsilon|z] = 0$, which is the orthogonality condition introduced in the NPIV model.[4]

We follow the control function approach as in NPV in order to deal with the endogeneity of $x$. Consider $E[y|x, z]$ which can be consistently estimated from data. Write

$$\begin{aligned} E[y|x, z] &= g_0(x, z_1) + E[\varepsilon|x, z] = g_0(x, z_1) + E[\varepsilon|\Pi_0(z) + v, z] \\ &= g_0(x, z_1) + E[\varepsilon|v, z] = g_0(x, z_1) + E[\varepsilon|v] \\ &= g_0(x, z_1) + \lambda_0(v) \qquad \text{(2.2)} \end{aligned}$$

where $\lambda_0(v) = E[\varepsilon|v]$, and the second last equality is from equation (2.1c). In effect, we capture endogeneity ($E[\varepsilon|x, z] \neq 0$) by an unobserved regressor $v$ which serves as a "control function." This is done so that the dependence structure between the two error terms (which is the source of endogeneity) is explicitly written as $\lambda_0(v)$.[5] Another intuition for this approach is that, with the endogenous variable $x = \Pi(z) + v$, once $v$ is controlled for or conditioned on, the only variation of $x$ comes from the exogenous variation of $z$. The control function approach has been introduced in Heckman (1979) for linear models, Smith and Blundell (1986), Rivers and Vuong (1988) for nonlinear models, and extended to nonparametric models by NPV with a separable control function, by Chesher (2003), Lee (2007) and Imbens and Newey (2009) with nonseparable control functions, and by Das, Newey and Vella (2004) with selection models, among others.[6]

Based on equation (2.2) we establish identification and estimation results.

[4] It is easy to show that if $v \perp z$ is assumed, then (2.1c) with $E[\varepsilon] = 0$ implies $E[\varepsilon|z] = 0$.

[5] A similar procedure can be found in Heckman (1979) where the selection bias component is captured by the dependence structure between the error terms, which is written as an inverse Mill's ratio.

[6] Note that the control function approach is sometimes called the "control variable" approach in the literature.

# 3 Identification

In this section, we obtain novel identification results which complement the identification results of NPV. Our results are also relevant to the subsequent weak identification analysis. We first restate the results of NPV for useful comparisons.

## 3.1 Results of Newey, Powell and Vella (1999)

Note that in (2.2), $g_0(x, z_1)$ is identified up to a constant if and only if, for any $\tilde{g}(x, z_1)$ and $\tilde{\lambda}(v)$ such that $E[y|x, z] = g_0(x, z_1) + \lambda_0(v) = \tilde{g}(x, z_1) + \tilde{\lambda}(v)$, we have that $\delta(x, z_1) = g_0(x, z_1) - \tilde{g}(x, z_1)$ is a constant function (and so is $\gamma(v) = \lambda_0(v) - \tilde{\lambda}(v)$). This idea motivates the following identification condition.

**Proposition 3.1 (Theorem 2.1 in NPV (p.565))** *$g_0(x, z_1)$ is identified up to an additive constant, if and only if $\Pr[\delta(x, z_1) + \gamma(v) = 0] = 1$ implies there is a constant $c_g$ with $\Pr[\delta(x, z_1) = c_g] = 1$.*

Identification of $g_0(x, z_1)$ is achieved if one can separately vary $(x, z_1)$ and $v$ in $g(x, z_1) + \lambda(v)$. Since $x = \Pi_0(z) + v$, a suitable condition on $\Pi_0(\cdot)$ will guarantee such separate variation of $x$ and $v$ via variation of $z$ and $v$. In light of this intuition, NPV propose an identification condition based on $\Pi(\cdot)$. Recall $z$ is partitioned as $z = (z_1, z_2)$.

**Proposition 3.2 (Theorem 2.3 in NPV (p.569))** *If $g(x, z_1)$, $\lambda(v)$, and $\Pi(z)$ are differentiable, the boundary of the support of $(z, v)$ has probability zero, and*

$$\Pr\left[rank\left(\frac{\partial \Pi_0(z)}{\partial z_2'}\right) = d_x\right] = 1, \tag{3.1}$$

*then $g_0(x, z_1)$ is identified.*

The identification condition can be seen as a nonparametric generalization of the rank condition. One can readily show that the order condition ($d_{z_2} \geq d_x$) is incorporated in this rank condition. Notice that the condition is only a sufficient condition, which suggests that the model can possibly be identified with a relaxed rank condition. This observation motivates our identification analysis.

## 3.2 Identification

We find a necessary and sufficient rank condition for identification by introducing a mild support condition. This analysis is important for the later purpose of defining the notion of *weak identification*. Given that the rank condition is necessary and sufficient, a “slight violation” of it has a binding effect on identification and hence results in the situation where the identification is weak. By introducing a localization technique, we can define weak identification by having the identification condition hold locally near the region of lack of identification.

Note that, as the rank condition of Proposition 3.2 is only a sufficient condition, a lack-of-identification condition cannot be derived from it. We first state and discuss the assumptions that we impose.

**Assumption ID1** *The functions $g(x, z_1)$, $\lambda(v)$, and $\Pi(z)$ are continuously differentiable in their arguments.*

This condition is also assumed in NPV; see Proposition 3.2 above. Before we state a key additional assumption for identification, we first define the supports of the random variables. Let $\mathcal{X} \subset \mathbb{R}^{d_x}$, $\mathcal{Z} \subset \mathbb{R}^{d_z}$, and $\mathcal{Z}_1 \subset \mathbb{R}^{d_{z_1}}$ be marginal supports of $x$, $z = (z_1, z_2)$, and $z_1$, respectively. Also, let $\mathcal{X}_z$ be the condition support of $x$ given $z \in \mathcal{Z}$. In identifying $g(x, z_1)$, it is useful to first conduct the analysis conditional on $z_1$, and then for all $z_1 \in \mathcal{Z}_1$. Let $\mathcal{X}_{z_1}$ and $\mathcal{Z}_{z_1}$ be the conditional support of $x$ given $z_1 \in \mathcal{Z}_1$ and the conditional support of $z$ given $z_1 \in \mathcal{Z}_1$, respectively. In order to incorporate a rank condition in the next identification assumption, we partition $\mathcal{Z}_{z_1}$ into two regions where the rank condition is satisfied and otherwise.

**Definition 3.3 (Relevant set)** *Given $z_1 \in \mathcal{Z}_1$, let $\mathcal{Z}_{z_1}^r$ be the subset of $\mathcal{Z}_{z_1}$ defined by*

$$\mathcal{Z}_{z_1}^r = \mathcal{Z}_{z_1}^r(\Pi_0(\cdot)) = \left\{ z \in \mathcal{Z}_{z_1} : rank\left(\frac{\partial \Pi_0(z)}{\partial z_2'}\right) = d_x \right\}.$$

The relevant set $\mathcal{Z}_{z_1}^r$ is where the instruments $z_2$ are relevant, so the identification power is determined by this set, which is discussed more precisely later. Also, let $\mathcal{Z}_{z_1}^0 = \mathcal{Z}_{z_1}^0(\Pi_0(\cdot)) = \mathcal{Z}_{z_1} \backslash \mathcal{Z}_{z_1}^r$ be the complement of the relevant set. In the univariate $x$ and $z_2$ case, $\mathcal{Z}_{z_1}^r$ is the region where $\Pi_0(\cdot)$ as a function of $z_2$ has nonzero slope and $\mathcal{Z}_{z_1}^0$ is the region where it is constant. Given $z_1 \in \mathcal{Z}_1$, let $\mathcal{X}_{z_1}^r$ be the subset of $\mathcal{X}_{z_1}$ defined by $\mathcal{X}_{z_1}^r = \left\{ x \in \mathcal{X}_z : z \in \mathcal{Z}_{z_1}^r \right\}$. Given these definitions, we introduce an additional support condition.

**Assumption ID2** *For any given $z_1 \in \mathcal{Z}_1$, the supports $\mathcal{X}_{z_1}$ and $\mathcal{X}_{z_1}^r$ differ only on a set of probability zero, i.e.,* $\Pr[x \in \mathcal{X}_{z_1} \backslash \mathcal{X}_{z_1}^r | z_1] = 0$ *almost surely.*

Intuitively, when $z_2$ is in the relevant set, $x = \Pi(z) + v$ varies as $z_2$ varies, and therefore the support of $x$ corresponding to the relevant set is large. Assumption ID2 assures that the corresponding support is large enough to almost surely cover the entire support of $x$. ID2 is not as strong as it may appear to be. Below, we provide mild sufficient conditions for ID2.

Next, although the support on which an unknown function is identified is usually left implicit, the following definition makes it more explicit in order to facilitate the proof.

**Definition 3.4 (Identification of a function)** *$g_0(x, z_1)$ is identified if $g_0(x, z_1)$ is identified on the support of $(x, z_1)$ almost surely.*

Note that this definition coincides with the identification concept implicitly assumed in Proposition 3.1. Suppose $z_1 \in \mathcal{Z}_1$ is fixed. Given the definition, if we identify $g_0(x, z_1)$ for any $x \in \mathcal{X}_{z_1}^r$, then we achieve identification of $g_0(x, z_1)$ by Assumption ID2. Now, in order to identify $g_0(x, z_1)$ for $x \in \mathcal{X}_{z_1}^r$, we need a rank condition, which is going to be minimal. The following is the result of identification:

**Theorem 3.5** *Suppose Assumptions ID1 and ID2 hold. Then $g_0(x, z_1)$ is identified up to an additive constant, if and only if,*

$$\Pr\left[ rank\left(\frac{\partial \Pi_0(z)}{\partial z_2'}\right) = d_x \middle| z_1 \right] > 0 \tag{3.2}$$

*almost surely.*

The proof is given after the following discussion.

The rank condition (3.2) is necessary and sufficient. By Definition 3.3, it can alternatively be written as $\Pr\left[z \in \mathcal{Z}_{z_1}^r | z_1\right] > 0$, a.s. The condition is substantially weaker than (3.1) in Proposition 3.2, since $\Pr\left[z \in \mathcal{Z}_{z_1}^r, z_1 \in \mathcal{Z}_1\right] = 1$ implies $\Pr\left[z \in \mathcal{Z}_{z_1}^r | z_1\right] = 1$ a.s. Conditional on $z_1$, it is enough for identification of $g_0(x, z_1)$ to have a (small) positive probability with which the rank condition is satisfied, which can be seen as the local rank condition as in Chesher (2003). That is, we achieve *global* identification with a *local* rank condition. This gain comes from having the additional support condition, but the trade-off is still appealing given the purpose of this identification analysis; later, we build a weak identification notion based on this necessary and sufficient rank condition.

Note that without Assumption ID2, we still achieve identification of $g_0(x, z_1)$ (up to a constant) under the assumptions of Theorem 3.5 but on the set $\left\{(x, z_1) : x \in \mathcal{X}_{z_1}^r, z_1 \in \mathcal{Z}_1\right\}$. Also, note that for identification of $g_0(x, z_1)$ at a given value of $z_1$, it is enough to have that (3.2) holds for such a value of $z_1$.

The following is a set of sufficient conditions that implies Assumption ID2. The proof is in the Appendix. Let $\mathcal{V}_z$ be the conditional support of $v$ given $z \in \mathcal{Z}$.

**Assumption ID2′** *The random variables $x$, $z_2$, and $v$ are continuously distributed and either* (a) *or* (b) *holds: For any given $z_1 \in \mathcal{Z}_1$,* **(a)** *(i) $x$ is univariate, (ii) $\mathcal{Z}_{z_1}$ is a cartesian product of connected intervals, and (iii) $\mathcal{V}_z = \mathcal{V}_{\tilde{z}}$ for all $z, \tilde{z} \in \mathcal{Z}_{z_1}^0$;* **(b)** *$\mathcal{V}_z = \mathbb{R}^{d_x}$, for all $z \in \mathcal{Z}_{z_1}$.*

The continuity of the r.v.'s is closely related to the support condition of Theorem 2.3 of NPV (Proposition 3.2) that the boundary of support of $(z, v)$ has probability zero. For example, when $z = (z_1, z_2)$ and $v$ are discrete their condition does not hold. Assumption ID2′(a)(i) assumes that the endogenous variable is univariate, which is most empirically relevant in nonparametric models. An additional condition is required for the multivariate $x$ case. Even under ID2′(a)(i), however, the exogenous covariates $z_1$ in $g(x, z_1)$ can still be a vector. ID2′(a)(ii) and (iii) are rather mild. ID2′(a)(ii) assumes that $z$ has a connected support conditional on $z_1$, which in turn requires that the instruments vary smoothly. ID2′(a)(iii) means that the conditional support of $v$ given $z$ is invariant when $z$ is in $\mathcal{Z}_{z_1}^0$. This *support invariance* condition is the key to have the considerably weaker rank condition in the identification theorem compared to that of NPV. Note that ID2′(a)(iii) along with the control function assumptions (2.1c)-(2.1d) is a more general set of assumptions for orthogonality of $z$ and $v$ than the full independence condition ($z \perp v$).

Note that $\mathcal{V}_z = \{x - \Pi_0(z) : x \in \mathcal{X}_z\}$. Therefore, ID2$'$(a)(iii) equivalently means that $\mathcal{X}_z$ is equivalent to $\mathcal{X}_{\tilde{z}}$ for $z$ and $\tilde{z}$ such that $E[x|z, z_2] = E[x|\tilde{z}, z_1] = const$. That is, the assumption implies that if a range of $x$ is realized with positive probability at a given $z$ then a similar range should be realized with positive probability at another given $\tilde{z}$ as long as $E[x|z_1, z_2]$ stays the same. With ID2$'$(a)(iii), the heteroskedasticity of $v$ which is previously allowed, may or may not be restricted. Note that this support invariance assumption can be tested from data.

Given ID2$'$(b) that the conditional support of $v$ is equal to $\mathbb{R}^{d_x}$, ID2 is trivially satisfied and no additional restriction is imposed on the joint support of $z$ and $v$. ID2$'$(b) also does not require univariate $x$ nor the connectedness of $\mathcal{Z}_{z_1}$. This assumption on $\mathcal{V}_z$ is satisfied with, for example, a normally distributed error term (conditional on regressors).

**Proof of Theorem 3.5:** The identification of $g_0(x, z_1)$ is achieved in two steps; first, we locally identify $g_0(x, z_1)$ in the sense of Chesher (2003), and then we achieve global identification. Consider equation (2.2), with $z = (z_1, z_2)$,

$$E[y|x, z] = E[y|v, z] = g_0(\Pi_0(z) + v, z_1) + \lambda_0(v), \tag{2.2}$$

and note that the conditional expectations and $\Pi_0(\cdot)$ are consistently estimable, and $v$ can also be estimated. By differentiating both sides of (2.2) with respect to $z_1$ and $z_2$, we have (for $d_x \times 1$ vectors $x$ and $v$, and a $d_z \times 1$ vector $z$)

$$\begin{aligned} \frac{\partial E[y|v, z]}{\partial z_1'} &= \frac{\partial g_0(x, z_1)}{\partial x'} \cdot \frac{\partial \Pi_0(z)}{\partial z_1'} + \frac{\partial g_0(x, z_1)}{\partial z_1'}, & (3.3) \\ \frac{\partial E[y|v, z]}{\partial z_2'} &= \frac{\partial g_0(x, z_1)}{\partial x'} \cdot \underbrace{\frac{\partial \Pi_0(z)}{\partial z_2'}}_{d_x \times d_{z_2}}. & (3.4) \end{aligned}$$

Now, for any fixed value $\bar{z}_1 \in \mathcal{Z}_1$, suppose $\Pr\left[z \in \mathcal{Z}^r_{\bar{z}_1} | z_1 = \bar{z}_1\right] > 0$. For any fixed value $\bar{z}_2$ such that $\bar{z} = (\bar{z}_1, \bar{z}_2) \in \mathcal{Z}^r_{\bar{z}_1}$, we have

$$rank\left(\frac{\partial \Pi_0(\bar{z})}{\partial z_2'}\right) = d_x, \tag{3.5}$$

by definition, hence the system of equations (3.4) has a unique solution $\frac{\partial g_0(x, \bar{z}_1)}{\partial x'}$ for $x \in \mathcal{X}_{\bar{z}}$. That is, $\frac{\partial g_0(x, \bar{z}_1)}{\partial x'}$ is locally identified for $x \in \mathcal{X}_{\bar{z}}$. (See Figure 1.) Note that due to the additive separability of the reduced-form error, the variation of $v$ in $\mathcal{V}$ sufficiently shifts the identifying location $x = \Pi_0(\bar{z}) + v$. See more discussion below on the separable structure of our problem.

The second part is to prove that the local rank condition is, indeed, enough for global identification of $g_0(x, z_1)$. Since the above argument is true for any $z = (\bar{z}_1, z_2) \in \mathcal{Z}^r_{\bar{z}_1}$, we have that $\frac{\partial g_0(x, \bar{z}_1)}{\partial x'}$ is identified on $x \in \mathcal{X}^r_{\bar{z}_1}$. (See Figure 1.) Now by Assumption ID2, the difference between $\mathcal{X}^r_{\bar{z}_1}$ and $\mathcal{X}_{\bar{z}_1}$ has probability zero. Thus $\frac{\partial g_0(x, \bar{z}_1)}{\partial x'}$ is identified by Definition 3.4.

Since $\Pr\left[z \in \mathcal{Z}^r_{z_1} | z_1\right] > 0$ a.s., the above argument is true for all $z_1 \in \mathcal{Z}_1$ a.s., and therefore

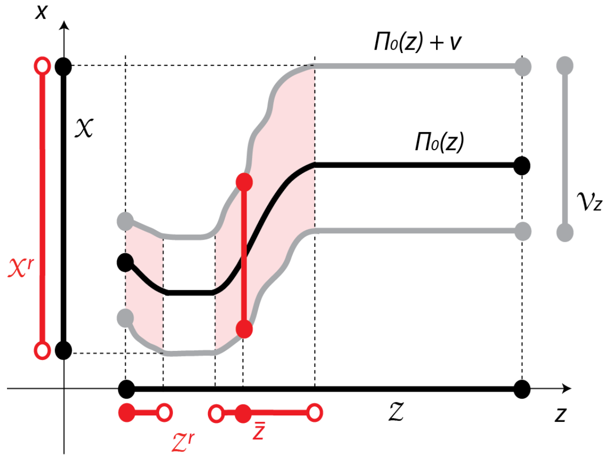

Figure 1: Identification under Assumption ID2$'$(a), univariate $z$ and no $z_1$.

$\frac{\partial g_0(x,z_1)}{\partial x'}$ is identified by Definition 3.4. This implies, $\frac{\partial g_0(x,z_1)}{\partial z_1'}$ is also identified by (3.3).[7] Consequently, $g_0(x, z_1)$ is identified up to an additive constant. The necessity part of the proof is in the Appendix. □

In order to identify the level of $g_0(x, z_1)$, we need to introduce some normalization as in NPV. Either $E[\varepsilon] = 0$ or $\lambda_0(\bar{v}) = \bar{\lambda}$ suffices to pin down $g_0(x, z_1)$. With the latter normalization, it follows $g_0(x, z_1) = E[y|x, z_1, v = \bar{v}] - \bar{\lambda}$, and we apply this normalization in estimation as it is convenient to implement.

Suppose $z = z_2$ is univariate and there is no $z_1$. Figure 1 illustrates the intuition of the identification proof under Assumption ID2$'$(a). With Assumption ID2$'$(b), the analysis is even more straightforward; see the proof of Lemma 11.2 in the Appendix. In the Figure, the local rank condition (3.2) ensures global identification of $g_0(x)$. The intuition of this identification result is the following. First, $g_0(x)$ is locally identified for $x$ corresponding to a point of $z$ in the relevant set $\mathcal{Z}^r$ by the rank condition. As such a point of $z$ varies within $\mathcal{Z}^r$, $x$ corresponding to it also varies enough to cover almost the entire support of $x$. At the same time, $x$ corresponding to irrelevant $z$ (i.e., $z$ outside of $\mathcal{Z}^r$) does not vary, while one can always find $z$ inside $\mathcal{Z}^r$ that gives the same value of such $x$.

When $\Pr[z \in \mathcal{Z}^r]$ is small but bounded away from zero, identification is still achieved, and the probability being small only affects the efficiency of estimators in the estimation stage. This issue is related to the weak identification concept discussed later; see Section 4.2.

Note that the strength of identification of $g_0(x)$ is different for different subsets of $\mathcal{X}$. For instance, identification must be strong in a subset of $\mathcal{X}$ corresponding to a subset of $\mathcal{Z}$ where $\Pi_0(\cdot)$ is steep. Or, over-identification can be present in a subset of $\mathcal{X}$ which corresponds to *multiple* subsets of $\mathcal{Z}$ where $\Pi_0(\cdot)$ has nonzero slope, so that multiple associations of $x$ and

[7] Once $\frac{\partial g_0(x,z_1)}{\partial x'}$ is identified, we can identify $\frac{\partial \lambda_0(v)}{\partial v'}$ by differentiating (2.2) w.r.t $v$:

$$\frac{\partial E[y|v,z]}{\partial v'} = \frac{\partial g_0(x,z_1)}{\partial x'} + \frac{\partial \lambda_0(v)}{\partial v}.$$

$z$ contribute to identification. This discussion implies that the shape of $\Pi_0(\cdot)$ provides useful information on the strength of identification in different parts of the domain of $g_0(x)$.

Lastly, it is worth mentioning that the separable structure of the reduced form along with ID2$'$(a)(iii) allows one to do global extrapolation in a manner that is analogous to global extrapolation in a linear model. If we had a linear model for the reduced form, then the local rank condition (3.2) would become a global rank condition and we would achieve global identification by Proposition 3.2. That is, this case is where linearity of the reduced-form function contributes to "global extrapolation" of the reduced-form relationship. Likewise, the identification results of this paper imply that although the reduced-form function is unrestricted, the way that the additive error interacts with other components of the model, such as the invariant support, enables global extrapolation of the relationship.

### 3.3 Rank Condition and Relevance of Instruments

Theorem 3.5 is useful in relating the *relevance of instruments* to the identification of $g_0(x, z_1)$. The nonparametric rank condition, i.e., $\Pr\left[rank\left(\partial\Pi_0(z)/\partial z_2'\right) = d_x|z_1\right] > 0$ a.s. implies restrictions on the shape of the conditional mean function $\Pi_0(z) = E[x|z_1, z_2]$ as a function of $z_2$. That is, Theorem 3.5 describes the degree of the nonlinear association of the instruments $z_2$ to the endogenous variables $x$ that is necessary and sufficient for identification. This is a more general way of considering the relevance of instruments than has been done in the literature with linear models, where the coefficients of instruments are the determinants of the relevance.

Ultimately, we are interested in how the weak relevance of instruments affects the performances of subsequent nonparametric estimators of $g_0(\cdot, \cdot)$, e.g., the asymptotic properties of series estimators. To facilitate the analysis, we consider the situation where (3.2) is "slightly violated." Since the condition is necessary and sufficient, a slight violation of it can effectively result in weak identification, and since the condition is related to the relevance of instruments, the concept of weak instruments is naturally defined. Note that this task will *not* be successful with (3.1), since violating the condition, i.e., $\Pr\left[rank\left(\partial\Pi_0(z)/\partial z_2'\right) = d_x\right] < 1$, can still result in the identification of the model.

To proceed with a more general setup, we establish a full range of strengths of instruments, i.e., irrelevant, weak, and strong instruments, in the framework of the lack of identification, weak identification and strong identification, respectively. The conditions for strong and irrelevant instruments are rather straightforward from the identification analysis above. The concept of weak instruments, or more generally, weak identification can be motivated by considering a situation where the "slope" of $\Pi_0(\cdot)$ is close to zero. This situation is discussed in detail in the next section. We start from the case of irrelevant instruments where the rank condition completely fails.

# 4 Lack of Identification and Weak Identification

## 4.1 Lack of Identification

The analysis of the lack of identification is important as a benchmark, and the case where the identification is weak can be constructed based on it. This is also a key approach Dufour (1997) takes in parametric models to formalize the concept of weak identification and address resulting inferential problems. Given the necessary and sufficient rank condition of Theorem 3.5, we can find the lack-of-identification condition, namely, the condition under which there exists $\tilde{g}(x, z_1) \neq g_0(x, z_1)$ (with positive probability) but is observationally equivalent to $g_0(x, z_1)$.

By contraposition of the necessity part of Theorem 3.5, we have the following theorem on lack of identification:

**Corollary 4.1** *Suppose Assumptions ID1 and ID2 hold. If*

$$\Pr\left[rank\left(\partial \Pi_0(z)/\partial z_2'\right) < d_x | z_1\right] = 1$$

*for some $z_1$ in a set with positive probability, then $g_0(x, z_1)$ is not identified, even up to an additive constant.*

Using the notation of the relevant set, this lack-of-identification condition is derived by negating the rank condition $\Pr[z \in \mathcal{Z}^r_{z_1} | z_1] \neq 0$ a.s., which gives $\Pr[z \in \mathcal{Z}^r_{z_1} | z_1] = 0$ for some $z_1$ with positive probability. The lack-of-identification condition is satisfied either when the number of the excluded instruments ($d_{z_2}$) is smaller than the number of the endogenous variables ($d_x$) so that the order condition fails, or when the excluded instruments are jointly irrelevant for one or more of the endogenous variables, almost everywhere in their support. With univariate $x$ and $z = z_2$, the condition simply becomes $\Pr[\partial \Pi_0(z)/\partial z = 0] = 1$, namely, that the function $\Pi_0(\cdot)$ is constant almost everywhere.

Let $\mathcal{C}(\mathcal{Z})$ be the class of conditional moment functions $\Pi(\cdot)$ on $\mathcal{Z}$ that are bounded, Lipschitz and continuously differentiable. Note that the derivative of $\Pi(\cdot)$ is bounded by the Lipschitz condition. We define a *non-identification region* $\mathcal{C}_0(\mathcal{Z})$ as a class of functions that satisfy the lack-of-identification condition:

$$\begin{aligned}\mathcal{C}_0(\mathcal{Z}) \quad = \quad & \{\Pi(\cdot) \in \mathcal{C}(\mathcal{Z}) : \Pr\left[rank\left(\partial \Pi_0(z)/\partial z_2'\right) < d_x | z_1\right] = 1, \\ & \qquad \text{for } z_1 \text{ in a set with positive probability}\}. \qquad (4.1)\end{aligned}$$

Also define $\mathcal{C}_1(\mathcal{Z}) = \mathcal{C}(\mathcal{Z}) \backslash \mathcal{C}_0(\mathcal{Z})$ as an *identification region.* With univariate $x$, $\mathcal{C}_0(\mathcal{Z})$ is merely an equivalence class of constant functions.

In the following subsection, we construct the notion of weak identification by a localization method around this non-identification region.

### 4.2 Weak Identification

In this section, we define *nonparametric weak instruments* as a sequence of reduced-form functions, which are localized around a function with no identification power. The sequence of models and localization technique are introduced to formally define weak instruments relative to the sample size $n$. As a result, the strength of instruments is represented in terms of a rate of the localization, and hence is eventually reflected in the convergence rate and asymptotic distribution (i.e., in local asymptotics) of the estimator of $g_0(\cdot,\cdot)$.[8]

We consider sequences of triangular models

$$y = g_0(x, z_1) + \varepsilon, \qquad x = \Pi_n(z) + v,$$

where $\Pi_n(\cdot)$ is sequences of functions in $\mathcal{C}_1(\mathcal{Z})$ (the identification region) which drift to a function $\bar{\Pi}(\cdot)$ in $\mathcal{C}_0(\mathcal{Z})$ (the non-identification region). The model is effectively localized around the non-identification region and the notion of weak identification thereby emerges. Although $g(x, z_1)$ is identified with $\Pi_n(\cdot) \in \mathcal{C}_1(\mathcal{Z})$ for any fixed $n$ by Theorem 3.5, as $\Pi_n(\cdot)$ drifts towards $\bar{\Pi}(\cdot)$, it can be said that $g(x, z_1)$ is only weakly identified. Intuitively, the function is weakly identified as the noise (i.e., $v$) contributes more than the signal (i.e., $\Pi_n(z)$) in the total variation of $x \in \{\Pi_n(z) + v : z \in \mathcal{Z}, v \in \mathcal{V}\}$. The following defines the notion of weak identification:

**Definition 4.2 (Weak identification)** *In the model (2.1), $g(x, z_1)$ is weakly identified (up to a constant) as $n \to \infty$, if (i) $\Pi_0(\cdot) = \Pi_n(\cdot) \in \mathcal{C}_1(\mathcal{Z})$ for $n \geq 1$, and (ii) there exists $\bar{\Pi}(\cdot) \in \mathcal{C}_0(\mathcal{Z})$ such that*

$$\left\| \Pi_n(z) - \bar{\Pi}(z) \right\| \to 0 \tag{4.2}$$

*almost surely, as $n \to \infty$.*

With univariate $x$ and $z = z_2$ for ease of exposition, condition (4.2) implies that $\Pi_n(\cdot)$ is modelled as "local to a constant function." This can also be seen as a result of "relaxing" the equality that appears in the lack-of-identification condition $\Pr[\partial\bar{\Pi}(z)/\partial z = 0] = 1$, that is,

$$|\partial \Pi_n(z)/\partial z| \to 0 \tag{4.3}$$

almost surely, as $n \to \infty$. Note that (4.3) characterizes the case where the slope of $\Pi_n(\cdot)$ becomes flatter as mentioned in Section 3.3. An example of the localization, that is, an example of nonparametric weak instruments, is depicted in Figure 2, which is also related to the next assumption.

In order to facilitate a meaningful asymptotic theory where the effect of weak identification (or weak instruments) is reflected, we further proceed by considering a specific sequence of

[8] The localization technique goes back to Pitman drift, which is used to analyze the local power properties of test statistics. In the weak instruments literature with linear models, e.g., Staiger and Stock (1997), drifting sequences of coefficients of instruments are used for the local asymptotic theory of IV estimators. Drifting sequences of functions are introduced e.g., in Stock and Wright (2000) with their parametric moment function in order to develop asymptotic theory for GMM estimators and test statistics under weak identification.

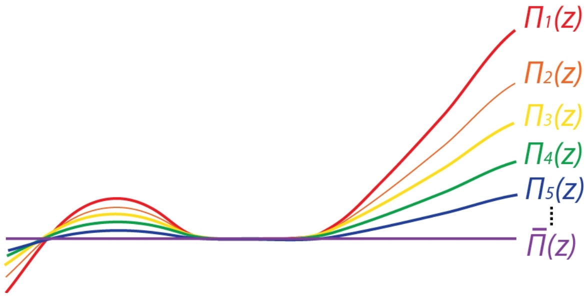

Figure 2: Nonparametric weak instruments by localization, univariate $x$ and $z$ and no $z_1$.

$\Pi_n(\cdot)$ with a certain rate. For vector $x$ and $z$, the following assumption defines nonparametric weak instruments as its special case by modeling the localization (4.3).

**Assumption L (Localization)** *For $\delta \in [0, \infty]$, $\Pi_0(\cdot) = \Pi_n(\cdot)$ satisfies the following. For some $\tilde{\Pi}(\cdot) \in \mathcal{C}_1(\mathcal{Z})$ that does not depend on $n$ and for $z \in \mathcal{Z}$*

$$\frac{\partial \Pi_n(z)}{\partial z_2'} = n^{-\delta} \cdot \frac{\partial \tilde{\Pi}(z)}{\partial z_2'} + o_p(n^{-\delta}).$$

Assumption L is equivalent to

$$\Pi_n(z) = n^{-\delta} \cdot \tilde{\Pi}(z) + c + o_p(n^{-\delta}) \tag{4.4}$$

for some constant vector $c$ by the differentiability of $\Pi(\cdot)$. Assumption L is a representation of a theoretical device called “local nesting” (Stock and Wright (2000)). Note that the sequence $\Pi_n(\cdot)$ in L uniformly converges to $\bar{\Pi}(\cdot)$ in $\mathcal{C}_0(\mathcal{Z})$.

Assumption L characterizes the relevance of instruments by embracing all categories of non-identification, weak identification and strong identification with corresponding values of $\delta$. Given Assumption L, the categories are summarized in Table 1; see Andrews and Cheng (2010, p. 4) for related discussions.

| Category | Sequence $n^{-\delta}$ | Identification |
|---|---|---|
| Strong IV | $\delta = 0$ | Strong ID of $g$ |
| Weak IV | $\delta \in (0, \infty)$ | Weak ID of $g$ |
| Irrelevant IV | $\delta = \infty$ | Lack of ID of $g$ |

Table 1: Identification Categories.

The parameter $\delta$ measures the strength of identification; the larger $\delta$ is, the weaker the instruments; when $\delta = 0$ we have the case of strong instruments.

The second category (weak instruments case) needs more discussion. In this category, Assumption L satisfies Definition 4.2. Given the discussion of weak identification above, the assumption provides a reasonable definition of *nonparametric weak instruments.* When $\delta \in (0, \infty)$, Assumption L implies (4.3) localization that the reduced form becomes flatter at the $n^\delta$ rate.[9]

The nonparametric version of weak instruments is harder to characterize than the "linear weak instruments." With a linear reduced form, weak instruments are specified simply with the coefficients of instruments, and the derivation of a local asymptotic theory is rather straightforward. With a nonparametric reduced form, on the other hand, we need to control the complete behavior of the reduced-form function and the derivation of local asymptotic theory seems to be more demanding. Nevertheless, Assumption L makes the weak instrument asymptotic theory straightforward, while it embraces the most interesting local alternatives (against the non-identification).

Moreover, with $\delta = 1/2$, the nonparametric weak instrument assumptions of Assumption L nest the linear weak instrument assumption in Staiger and Stock (1997). In linear case where $\Pi(z) = z_2'\pi$, we have $\partial\Pi(z)/\partial z_2' = \pi$. By letting $\partial\tilde{\Pi}(z)/\partial z_2' = \Delta$ for some fixed constant $\Delta$ and $\delta = 1/2$, Assumption L coincides with the weak instrument specification ($\pi = \Delta/\sqrt{n}$) of Assumption $\mathrm{L}_\Pi$ in Staiger and Stock (1997, p. 560). Therefore, Assumption L for $\delta \in (0, \infty)$ can be seen as the nonparametric generalization of the weak instruments (or, local to zero modeling) in linear simultaneous equations models in the literature.

# 5 Estimation

Before we develop the main estimation procedure of this paper, we revisit the existing series estimation procedure in the literature which uses the control function approach. Then we characterize the problem of weak instruments in the procedure as a multicollinearity problem, and also relate it to the ill-posed inverse problem. This relationship motivates the introduction of penalization as a regularization method for weak instruments in the series estimation procedure, which leads to penalized series estimation.

## 5.1 Control Function Approach and Existing Estimation Method

In this subsection, with model (2.1), we review the standard series estimator established in NPV. Henceforth, in order to keep our presentation succinct, we focus on the case where the included exogenous variables $z_1$ is dropped from model (2.1). With $z_1$ included, the estimation analysis follows along similar lines. Then, equation (2.2) becomes

$$E[y|x, z] = g_0(x) + \lambda_0(v). \tag{2.2'}$$

[9] It would be interesting to have different rates across columns or rows of $\frac{\partial\tilde{\Pi}(\cdot)}{\partial z_2'}$. One can also consider different rates for different elements of the matrix. The analyses in these cases can analogously be done by slight motifications of the arguments.

Note that equation (2.2′) can be rewritten as

$$y = g_0(x) + \lambda_0(v) + \eta = h_0(w) + \eta \tag{5.1}$$

where $w = (x, v)$ and $\eta = y - E[y|x, z]$ so that $E[\eta|x, z] = 0$ by definition. Once the endogeneity is controlled by the control function, the problem becomes one of estimating the additive nonparametric regression function $h_0(w)$. Since the reduced-form error $v$ is unobserved, the procedure takes two steps. In the first stage, we estimate the reduced form $\Pi_0(\cdot)$ and obtain the residual $\hat{v}$. In the second stage, we estimate structural function $h_0(\cdot)$ with $\hat{w} = (x, \hat{v})$ as regressors, where $\hat{v}$ is a generated regressor. We consider series estimation for both $\Pi_0(\cdot)$ and $h_0(\cdot)$. With this method, it is easy to impose the additivity of $h_0(\cdot)$ and also to characterize the problem of weak instruments. The main conclusions of this paper, however, do not depend on the choice of estimation method; see Section 7 for a related discussion.

Let $\{(y_i, x_i, z_i)\}_{i=1}^n$ be the data with $n$ observations, and let $r^L(z_i) = (r_1(z_i), ..., r_L(z_i))'$ be a vector of approximating functions (e.g. polynomials or splines) of order $L$ for the first stage.[10] Define a matrix $\underset{n\times L}{R} = (r^L(z_1), ..., r^L(z_n))'$. Then regressing $x_i$ on $r^L(z_i)$ gives

$$\hat{\Pi}(\cdot) = r^L(\cdot)'\hat{\gamma}, \qquad \hat{\gamma} = (R'R)^{-1}R'(x_1, ..., x_n)'. \tag{5.2}$$

Now, we use the residuals, $\hat{v}_i = x_i - \hat{\Pi}(z_i)$, as a control function in the second stage. Define a vector of approximating functions of orders $K = K_1 + K_2 - 1$ for the second stage, where $K_1 \geq 2$ and $K_2 \geq 2$,

$$p^K(w) = (1, p_2(x), ..., p_{K_1}(x), p_2(v), ..., p_{K_2}(v))' = \left[1 \vdots p_-^{K_1}(x)' \vdots p_-^{K_2}(v)'\right]'.$$

Note that the subvectors $p_-^{K_1}(x)$ and $p_-^{K_2}(v)$ are vectors of approximating functions for $g_0(\cdot)$ and $\lambda_0(\cdot)$ of orders $K_1 - 1$ and $K_2 - 1$, respectively, where the first elements of $p^{K_1}(x)$ and $p^{K_2}(v)$, i.e., $p_1(x) = p_1(v) = 1$ (both in the polynomial and spline cases), are dropped. Since $g_0(\cdot)$ and $\lambda_0(\cdot)$ can only be separately identified up to a constant, when estimating $h_0(\cdot)$, we include only one constant function. Also, to reflect the additive structure of (2.2′), there are no interaction terms in the vector. Consider a matrix of approximating functions:

$$\begin{aligned} \underset{n\times K}{\hat{P}} &= (p^K(\hat{w}_1), ..., p^K(\hat{w}_n))' \\ &= \begin{bmatrix} 1 & p_2(x_1) & \cdots & p_{K_1}(x_1) & p_2(\hat{v}_1) & \cdots & p_{K_2}(\hat{v}_1) \\ \vdots & & & & & & \vdots \\ 1 & p_2(x_n) & \cdots & p_{K_1}(x_n) & p_2(\hat{v}_n) & \cdots & p_{K_2}(\hat{v}_n) \end{bmatrix}. \end{aligned} \tag{5.3}$$

Then, the series estimator is obtained by regressing $y_i$ on $p^K(\hat{w}_i)$:

$$\hat{h}(\cdot) = p^K(\cdot)'\hat{\beta}, \qquad \hat{\beta} = (\hat{P}'\hat{P})^{-1}\hat{P}'Y, \tag{5.4}$$

[10] For detailed descriptions of power series and regression splines for $r^L(z)$ and $p^K(w)$, see pp. 572-573 in NPV.

where $Y = (y_1, ..., y_n)'$. In fact, $\hat{\beta}$ and $\hat{\gamma}$ are classical least square estimators if their dimensions $K$ and $L$, respectively, are fixed. Here, however, it is important to notice that $L = L(n)$, $K_1 = K_1(n)$ and $K_2 = K_2(n)$ (hence $K = K(n)$) grow with the sample size $n$, which implies that the size of the matrix $\hat{P}$ as well as $R$ also grows with $n$. It is important to account for this aspect in deriving asymptotics results. The standard asymptotics results of series estimators (without the consideration of weak instruments) can be found in Andrews (1991), Newey (1997), and NPV.

Given the estimator $\hat{h}(\cdot)$, with the normalization that $\lambda(\bar{v}) = \bar{\lambda}$, we have

$$\hat{g}(x) = \hat{h}(x, \bar{v}) - \bar{\lambda}. \tag{5.5}$$

### 5.2 Weak Instruments, Multicollinearity, and the Ill-Posed Inverse Problem

In a weak instrument environment, we face a nonstandard problem in estimating $h_0(w) = g_0(x) + \lambda_0(v)$ using the procedure discussed above. To facilitate the discussion, we consider a series representation of the nonparametric regression equation (5.1):

$$y = g_0(x) + \lambda_0(v) + \eta = \sum_{j=1}^{\infty} \left\{\beta_{1j} p_j(x) + \beta_{2j} p_j(v)\right\} + \eta, \tag{5.6}$$

where the $p_j(\cdot)$'s are the approximating functions (or basis functions). Note that under the weak instrument specification (4.4) of Assumption L in Section 4.2, as $n \to \infty$, $E[x|z] = \Pi_n(z) \to c$. If $c = 0$, then

$$v = x - \Pi_n(z) \to x \quad \text{a.s.}$$

(or more precisely, $|x - v| = O_p(n^{-\delta})$) and $p_j(v) \to p_j(x)$ a.s., for all $2 \leq j \leq K(n)$. That is, weak instruments result in the problem where the two regressors are nearly identical in (5.6). If $c \neq 0$, then $x$ and $v$ just differ by a fixed constant.[11]

This feature is manifested in the series estimation as a multicollinearity problem. Given that $\hat{v}$ and $v$ are "close" to each other, two columns of the regressor matrix $\hat{P}$ in (5.3) become nearly identical. This problem corresponds to so called *near multicollinearity*.[12]

[11] Jiang, Fan and Fan (2010) precisely consider this instance of correlated regressors in a simple nonparametric additive model, where the correlation is caused by the nature of dataset they concern; see Section 7 of the present paper for more discussion.

[12] See Hastie and Tibshirani (1990) for the discussion of *concurvity*, a nonlinear/nonparametric analogue of multicollinearity.

In fact, the same argument can also be made in a linear simultaneous equations model analogous to the present triangular model once we apply a similar control function approach there. First note that the control function estimator is equivalent to the usual two stage least squares (TSLS) estimator in linear models. Therefore, the problem of weak instruments with the TSLS estimator analyzed in the literature such as Staiger and Stock (1997) can be seen as the multicollinearity problem with the control function estimator. In such linear settings, however, introduction of a regularization method discussed in this paper is not well-justified; refer to further discussion in the current subsection.

Define a $K \times K$ sample second moment matrix

$$\hat{Q} = \frac{\hat{P}'\hat{P}}{n} = \frac{\sum_{i=1}^{n} p^K(\hat{w}_i)p^K(\hat{w}_i)'}{n}.$$

Under Assumption L, matrix $\hat{Q}$ becomes nearly singular or "local to singular" due to the multicollinearity in $\hat{P}$, which is problematic when inverting the matrix to calculate $\hat{\beta}$ in (5.4). Consequently, under the weak instrument assumption the performance of $\hat{h}(\cdot)$ deteriorates severely. In Section 6.3 below, we introduce a regularization method for estimating $h_0(\cdot)$ to improve the performance of the resulting estimator, which is done by controlling the singularity problem of the sample second moment matrix.

To motivate such a regularization scheme, for the remainder of this subsection, we compare and also contrast the problem of weak instruments with the ill-posed inverse problem in the literature with illustrative examples. In doing so, we justify which of the regularization methods in the literature properly works in the problem of weak instruments.

The ill-posed inverse problem is a function space inversion problem that typically occurs in a standard NPIV approach. Consider the model

$$y = g_0(x) + \varepsilon, \quad E[\varepsilon|z] = 0$$

which implies

$$E[y|z] = E[g_0(x)|z] = \int g_0(x)dF(x|z). \tag{5.7}$$

Equation (5.7) is an Fredholm integral equation of the first kind, where the inverse problem of recovering $g_0(\cdot)$ is ill-posed because the estimated observable (the reduced form $E[y|z]$) has a discontinuous effect on the object we recover. An illustration of the ill-posed inverse problem in this NPIV approach can be found, e.g., in Horowitz (2011).

**Example 5.1—The Ill-Posed Inverse Problem in the NPIV Approach:** Following the example of Horowitz (2011), let $f(x,z) = \sum_{j=1}^{\infty} \lambda_j^{1/2}\phi_j(x)\phi_j(z)$, where $\lambda_j$'s are eigenvalues of an integral operator produced from (5.7). Also let $x$ and $z$ be uniformly distributed on $[0,1]$. Then, with $\tilde{\eta} = y - E[y|z]$ and the generalized Fourier coefficient $\gamma_j$

$$y = E[g_0(x)|z] + \tilde{\eta} = \sum_{j=1}^{\infty} \gamma_j E[\phi_j(x)|z] + \tilde{\eta} = \sum_{j=1}^{\infty} \gamma_j \lambda_j^{1/2}\phi_j(z) + \tilde{\eta}, \tag{5.8}$$

where the last equality is due to $E[\phi_j(x)|z] = \int_0^1 \phi_j(x)f_{x|z}(x|z)dx = \int_0^1 \phi_j(x)f_{x,z}(x,z)dx = \lambda_j^{1/2}\phi_j(z)$. Here, the ill-posed inverse problem arises because the $E[\phi_j(x)|z]$'s do not have much variation even though the basis functions $\phi_j(x)$'s do (Newey and Powell (2003, p. 1568)), or because $\lambda_j^{1/2}\phi_j(\cdot) \to 0$ by the fact that $\lambda_j \to 0$ as $j \to \infty$. This problem can alternatively be seen by letting $c_j = \gamma_j\lambda_j^{1/2}$ be the coefficient of the regressor $\phi_j(x)$ and $\hat{c}_j$ the resulting estimator. Then the estimator of $\gamma_j$ is $\hat{c}_j/\lambda_j^{1/2}$, where the denominator converges to zero as $j \to \infty$. This implies instability of estimators of $\gamma_j$, which prevents the accurate estimation of $g_0(\cdot)$.

In order to tackle the ill-posed inverse problem, two approaches are taken in the NPIV literature: *the truncation method* and *the penalization method*.[13] Using the example above, the truncation method is to regularize the problem by replacing (5.8) with a finite-dimensional approximation. This is done by truncating the infinite sum, that is, by considering $j \leq J_n$ for some $J_n < \infty$ for all $n$. In this way, one prevents the $\lambda_j$'s from converging to zero as $j \to \infty$. The penalization method is to directly control the behavior of coefficients $\gamma_j$'s for all $j < \infty$ by penalizing them, while maintaining the original infinite-dimensional approximation. For more discussion, see Chen and Pouzo (2009). □

The weak instrument problem in the triangular model of the present paper has a similar feature to the ill-posed inverse problem above. It is, however, important to notice that the nature of the problem is slightly different and penalization but not truncation works in regularizing the weak instrument problem. To see this, we consider an example where the reduced-form equation is simplified for ease of exposition.

**Example 5.2—The Weak instrument Problem in the Control Function Approach:** Let $x = \pi_n z + v$ be the reduced-form equation with univariate $x$ and $z$ and but otherwise the model is the same as in (2.1). Note that the instrument is weak because $\pi_n \to 0$ as $n \to \infty$. By using the mean value expansion, it follows $p_j(v) = p_j(x - \pi_n z) = p_j(x) - \pi_n z p_j'(\tilde{x})$ where $\tilde{x}$ is an intermediate value. Then, by plugging this in the expression in (5.6) and rearranging terms, it follows that

$$y = \sum_{j=1}^{\infty} \left\{ (\beta_{1j} + \beta_{2j}) p_j(x) - \beta_{2j} \pi_n z \partial p_j(\tilde{x})/\partial x \right\} + \eta. \tag{5.4$'$}$$

Here, the weak instrument problem arises because, for all $j$, the $\pi_n z \partial p_j(\tilde{x})/\partial x$ terms converge to zero and hence their variation shrinks, as $n \to \infty$. Alternatively, let $d_j = \beta_{2j}\pi_n$ be the coefficient on the regressor $z\partial p_j(\tilde{x})/\partial x$ and $\hat{d}_j$ the resulting estimator. Then the estimator of $\beta_{2j}$ is $\hat{d}_j/\pi_n$, where the denominator converges to zero as $n \to \infty$. This implies that the estimator of $\beta_{2j}$ is unstable, and so is the estimator of $\beta_{1j}$ which is recovered from the estimator of $\beta_{1j} + \beta_{2j}$. As a regularization method to tackle the weak instrument (or multicollinearity) problem, the truncation method does not work properly. Unlike (5.8) in the ill-posed inverse problem, the estimator of $\beta_{2j}$ can still be unstable even after truncating the series since, for $j \leq J < \infty$, we still have $\beta_{2j} = d_j/\pi_n \to \infty$ as $n \to \infty$. On the other hand, the penalization directly controls the behavior of $\beta_{2j}$'s, and hence successfully regularizes the weak instrument problem. □

The discussion of this subsection is summarized for convenience in Figure 3. The numbering in the diagram corresponds to the numbering in the remarks below. The three concepts

[13] In Chen and Pouzo (2009), closely related concepts are considered with different terminologies: minimizing a criterion over finite sieve space and minimizing a criterion over infinite sieve space with Tikhonov-type penalty, respectively.

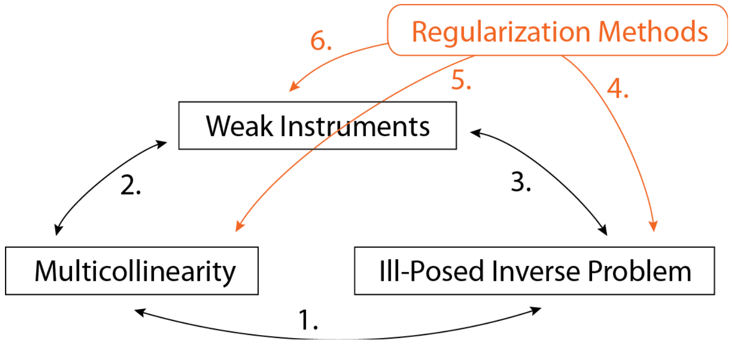

Figure 3: Weak instruments, multicollinearity, and the ill-posed inverse problem.

of *the problem of weak instruments*, *ill-posed inverse problem*, and *multicollinearity* are interrelated:

1. The ill-posed inverse problem can be seen as a multicollinearity problem. "[T]he ill-posed inverse problem is a functional analogue to the problem of multicollinearity in a classical linear regression model, where large differences in regression coefficients can correspond to small differences in fitted values of the regression function." (Blundell and Powell (2006, p. 321))

2. We find that the weak instrument problem can be viewed as a multicollinearity problem in the control function framework.

3. The weak instrument and ill-posed inverse problems are also related to each other, which is implied by the connections of 1 and 2. Examples 5.1-5.2 above also show their similarities, while drawing clear distinctions between the two.

Given the discussion, strategies that can tackle those problems must also be interrelated:

4. In the NPIV literature, regularization methods are introduced to deal with the ill-posed inverse problem in estimation; see, e.g., Newey and Powell (2003), Ai and Chen (2003), Blundell, Chen and Kristensen (2007), Hall and Horowitz (2005), Horowitz and Lee (2007), and Chen and Pouzo (2009). Among others, Chen and Pouzo (2009) introduce the penalized sieve minimum distance estimator, which essentially incorporates both the truncation and penalization methods.

5. The standard regularization approach for multicollinearity is a biased estimation method called the ridge regression; see Hoerl and Kennard (1970). The method can be seen as a penalization method.

6. Given the connections of 1-3, regularization methods used in the realm of research concerning inverse problems as in 4 and 5 are suitable for use with weak instruments. Among the two regularization methods, the penalization scheme is introduced in this paper. As discussed in Examples 5.1-5.2, unlike the truncation method, the penalization method alleviates the multicollinearity problem induced by weak instruments. It is not surprising that the resulting penalized series estimator has a ridge regression form of 5; see below.

## 5.3 Penalized Series Estimation

Given the discussion in the previous subsection, we introduce a penalization scheme to alleviate the weak instrument effect in the original series estimator of $h_0(\cdot)$ defined in (5.2)-(5.5).

We define a *penalized series estimator*:

$$\hat{h}_\tau(w) = p^K(w)'\hat{\beta}_\tau, \tag{5.9}$$

where the "interim" estimator $\hat{\beta}_\tau$ optimizes a penalizing criterion function,

$$\hat{\beta}_\tau = \arg\min_\beta \left(y - \hat{P}\beta\right)' \left(y - \hat{P}\beta\right)/n + \tau_n \beta'\beta, \tag{5.10}$$

where $\tau_n \geq 0$ is the penalization parameter that satisfies $\tau_n \to 0$ as $n \to \infty$ and $\hat{P}$ is defined in (5.3). The penalty term $\tau_n\beta'\beta$ penalizes the objective function more when $\|\beta\|$ is large, which effectively imposes restrictions on $h_0(\cdot)$ by imposing $\|\beta\| \leq B$ for some fixed constant $B < \infty$. In fact, the method of penalization incorporates prior information on the true structural function such as smoothness properties (Chen and Pouzo (2009)). In most cases of econometric modeling, the structural function derived from economic models cannot be too "wiggly." This is also the rationale of imposing smoothness assumptions in various nonparametric estimators. The smoothness assumptions are also related to the regularization methods discussed in the previous section. As previously discussed, however, incorporating the smoothness assumption specifically through the penalization alleviates the weak instrument effect. Let $\beta_- = (\beta_2, \beta_3, ..., \beta_K)'$. Then, $\|\beta\| \leq B$ above implies that $\left\|\beta_-\right\| \leq B$ and $|\beta_1| \leq B$, which are related to the smoothness and bounded intercept of $h_0(\cdot)$, respectively. To only ensure smoothness, we can penalize the higher terms of $\beta$, i.e., $\beta_-$ and similar arguments go through.

We have a closed form solution for the optimization problem (5.10):

$$\hat{\beta}_\tau = (\hat{P}'\hat{P} + n\tau_n I)^{-1}\hat{P}'y = (\hat{Q} + \tau_n I)^{-1}\hat{P}'y/n.$$

As with the ridge regression estimator, the term $\tau_n I$ mitigates the singularity of the matrix $\hat{Q}$ caused by weak instruments or multicollinearity.

This penalized estimation method is appealing in the present nonparametric settings. Unlike the biased estimator of ridge regression in a linear model, we do not interpret $\hat{\beta}_\tau$ here, since it is only an interim estimator used to obtain $\hat{h}_\tau(\cdot)$. Moreover, the overall bias of $\hat{h}_\tau(\cdot)$ is unlikely to be worsened in the sense that the additional bias introduced by penalization can be dominated by the existing series estimation bias. Indeed, the penalized series estimator is shown to be consistent provided that the instruments are not severely weak and the penalization parameter shrinks fast enough; see below.

### 5.4 Weak Instrument Effect and Penalization Effect

This subsection outlines key technical steps of this paper that are useful in deriving the local asymptotic properties (i.e., the convergence rate and asymptotic normality) of the penalized series estimator $\hat{h}_\tau(\cdot)$. The steps are only theoretical procedures. The performance of $\hat{h}_\tau(\cdot)$ under the weak instrument environment can be previewed along the way. Let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the maximum and minimum eigenvalues of a symmetric matrix $A$, respectively.

In $\hat{\beta}_\tau = (\hat{Q}+\tau_n I)^{-1}\hat{P}'y/n$, $\hat{Q}$ is nearly singular and $\tau_n I$ is the additional penalization term controlling it. Let $\hat{Q}_\tau = \hat{Q} + \tau_n I$. Note that the "degree of singularity" of $\hat{Q}_\tau$ is important in the asymptotic performance of $\hat{h}_\tau(\cdot)$, and it is determined by the relative effect of weak instruments and penalization. We establish this claim by calculating the minimum eigenvalue of $\hat{Q}_\tau$, or equivalently, maximum eigenvalue of $\hat{Q}_\tau^{-1}$.[14] First, note that, for $\tau_n > 0$ and using $\lambda_{\min}(\hat{Q})^{-1} = \lambda_{\max}(\hat{Q}^{-1})$, we have

$$\lambda_{\max}(\hat{Q}_\tau^{-1}) = \frac{1}{\lambda_{\min}(\hat{Q}+\tau_n I)} \le \frac{1}{\lambda_{\min}(\hat{Q}) + \lambda_{\min}(\tau_n I)} \le \min\left\{\lambda_{\max}(\hat{Q}^{-1}), \tau_n^{-1}\right\}. \tag{5.11}$$

For the first inequality, see Lemma 11.4 in the Appendix. Note that each term inside the minimum function goes to infinity at a particular rate as $n \to \infty$. We bound $\lambda_{\max}(\hat{Q}^{-1})$ by a quantity that depends on the $n^\delta$ rate of Assumption L by linearizing the approximating functions in $\hat{Q}$. As the population version of $\hat{Q}$ has the same feature as $\hat{Q}$, by defining the population second moment matrix

$$Q = E[p^K(w_i)p^K(w_i)'], \tag{5.12}$$

for ease of exposition, we calculate the order of magnitude of $\lambda_{\max}(Q^{-1})$ instead. The Appendix gives bounds for both matrices. Under Assumption L, due to the singularity discussed above, we do *not* have a standard condition such as Assumption A1 in NPV (p.593) that $\lambda_{\min}(Q)$ is bounded away from zero uniformly over $K(n)$, i.e., $\lambda_{\min}(Q) \ge c > 0$, or equivalently $\lambda_{\max}\left(Q^{-1}\right) \le C < \infty$, for some constants $c$ and $C$ that do not depend on $n$ and uniformly for all $K(n)$. Deriving the rate at which $\lambda_{\max}\left(Q^{-1}\right)$ diverges provides insight into the extent to which the condition $\lambda_{\min}(Q) \ge c$ is violated.

For the rest of this subsection, we consider univariate $x$ for simplicity.[15] Note that $z$ is still a vector, and as above we omit $z_1$ in the structural function. Impose Assumption L, then $\Pi_n(\cdot) = n^{-\delta}\tilde{\Pi}(\cdot)$ by (4.4) after applying a normalization that $c = 0$ and suppressing $o_p(n^{-\delta})$ for simplicity. Omitting $o_p(n^{-\delta})$ does not affect the asymptotic results developed in the paper. Assume that the approximating function $p_j(x_i)$ is twice differentiable for all $j$, and for $r \in \{1, 2\}$, define its $r$th derivative as $\partial^r p_j(x) = d^r p_j(x)/dx^r$.

[14] We consider singularity or invertibility of the matrix in terms of its eigenvalues as the order $K(n) \times K(n)$ of this matrix is growing with $n$. This approach is standard in the series estimation literature (e.g., Andrews (1991), Newey (1997), and NPV) or sieve estimation literature (e.g., Blundell, Chen and Kristensen (2007) and Chen and Pouzo (2009)) to impose conditions on second moment matrices for asymptotic theory.

[15] The analysis can also be generalized to the case of a vector $x$ by using multivariate mean value expansion. See the Appendix for discussion of this generalization.

By mean value expanding each element of $p^{K_1}(x_i)$ around $v_i$, we have, for $2 \leq j \leq K_1$,

$$p_j(x_i) = p_j(n^{-\delta}\tilde{\Pi}(z_i) + v_i) = p_j(v_i) + n^{-\delta}\tilde{\Pi}(z_i)\partial p_j(\tilde{v}_i), \tag{5.13}$$

where $\tilde{v}_i$ is a value between $x_i$ and $v_i$. Define $\partial^r p_{-}^{K_1}(x) = [\partial^r p_2(x), \partial^r p_3(x), ..., \partial^r p_{K_1}(x)]'$. Then by (5.13) the vector of regressors $p^K(w_i)$ for estimating $h(\cdot)$ can be written as

$$p^K(w_i)' = \left[1 \vdots p_{-}^{K_1}(x_i)' \vdots p_{-}^{K_2}(v_i)'\right] = \left[1 \vdots p_{-}^{K_1}(v_i)' + n^{-\delta}\tilde{\Pi}(z_i)\partial p_{-}^{K_1}(\tilde{v}_i)' \vdots p_{-}^{K_2}(v_i)'\right]. \tag{5.14}$$

For expositional convenience, we assume the vectors of regressors for $g_0(\cdot)$ and $\lambda_0(\cdot)$ have the same orders $\kappa$, i.e., $\kappa = \kappa(n) = K_1 = K_2 = (K+1)/2$. The discussion for the general case can be found in the Appendix.

Now we choose a transformation matrix $T_n$ to be

$$T_n = \begin{bmatrix} 1 & 0_{1\times\kappa} & 0_{1\times\kappa} \\ 0_{\kappa\times 1} & n^{\delta} I_\kappa & 0_{\kappa\times\kappa} \\ 0_{\kappa\times 1} & -n^{\delta} I_\kappa & I_\kappa \end{bmatrix}.$$

so that after multiplying both sides of (5.14) by $T_n$, the weak instrument factor is removed from $p^K(w_i)'$. That is, we have

$$\begin{aligned} p^K(w_i)'T_n &= [1 \vdots p_{-}^{\kappa}(v_i)' + n^{-\delta}\tilde{\Pi}(z_i)\partial p_{-}^{\kappa}(\tilde{v}_i)' \vdots p_{-}^{\kappa}(v_i)'] \cdot T_n \\ &= [1 \vdots \tilde{\Pi}(z_i)\partial p_{-}^{\kappa}(\tilde{v}_i)' \vdots p_{-}^{\kappa}(v_i)'] \\ &= [1 \vdots \tilde{\Pi}(z_i)\partial p_{-}^{\kappa}(v_i)' \vdots p_{-}^{\kappa}(v_i)'] + [0 \vdots \tilde{\Pi}(z_i)\left(\partial p_{-}^{\kappa}(\tilde{v}_i)' - \partial p_{-}^{\kappa}(v_i)'\right) \vdots (0_{\kappa\times 1})'] \\ &= p^{*K}(u_i)' + m_i^{K\prime} \end{aligned} \tag{5.15}$$

where $u_i = (z_i, v_i)$ and $p^{*K}(u_i)$ and $m_i^K$ are defined implicitly. To illustrate the role of this linear transformation, rewrite the original vector of regressors in (5.14) as

$$p^K(w_i)' = p^K(w_i)'T_n T_n^{-1} = \left\{p^{*K}(u_i) + m_i^K\right\}' T_n^{-1}.$$

Ignoring the remainder vector $m_i^K$ which is shown to be asymptotically negligible in the proof, the original vector $p^K(w_i)$ is separated into two components, namely, $p^{*K}(u_i)$ and $T_n^{-1}$. Note that $p^{*K}(u_i) = [1 \vdots \tilde{\Pi}(z_i)\partial p_{-}^{\kappa}(v_i)' \vdots p_{-}^{\kappa}(v_i)']'$ is not affected by the weak instruments and can be seen as a new set of regressors.[16] Also note that some calculations in the Appendix show that $\lambda_{\min}(T_n^{-1}) = O(n^{-\delta})$, which captures the weak instrument effect of Assumption L.

By equations (5.12) and (5.15), it follows

$$\begin{aligned} T_n' Q T_n &= E[T_n' p^K(w_i) p^K(w_i)' T_n] \\ &= Q^* + E\left[m_i^K p^{*K}(u_i)'\right] + E\left[p^{*K}(u_i) m_i^{K\prime}\right] + E\left[m_i^K m_i^{K\prime}\right] \end{aligned} \tag{5.16}$$

[16] For the justification that $p^{*K}(u_i)$ can be seen as a vector of regressors, see Assumption B in Section 6.1 and Assumption B.1 in Section 11.2 of Appendix.

where the newly defined $Q^* = E\left[p^{*K}(u_i)p^{*K}(u_i)'\right]$ is the population second moment matrix of the new regressors. Since $Q^*$ is not affected by the weak instruments, we can safely assume that $\lambda_{\min}(Q^*) \geq c > 0$, or equivalently, $\lambda_{\max}\left(Q^{*-1}\right) \leq C < \infty$, both uniformly in $K(n)$ (Assumption B.1 in the Appendix). Since the remaining three terms in equation (5.16) can be shown to be asymptotically negligible, this implies that $\lambda_{\max}((T_n'QT_n)^{-1}) \leq C < \infty$ uniformly in $K(n)$. Therefore, it follows that

$$\lambda_{\max}(Q^{-1}) = \lambda_{\max}(T_n(T_n'QT_n)^{-1}T_n') \leq \lambda_{\max}((T_n'QT_n)^{-1})\lambda_{\max}(T_nT_n') \leq O(1)O(n^{2\delta}),$$

and the order of magnitude of $\lambda_{\max}(Q^{-1})$ is found to be $n^{2\delta}$. This shows the degree of singularity of $Q$ in terms of the weak instrument rate specified in Assumption L. Note that when the instruments are strong (i.e., $\delta = 0$), $Q^{-1}$ is bounded (i.e., $\lambda_{\max}(Q^{-1}) = O(1)$) and the usual condition in the literature can be imposed. The rigorous derivation of this conclusion, as well as a similar derivation with $\hat{Q}$ can be found in the Appendix. Using these results and (5.11), we obtain

$$\lambda_{\max}(\hat{Q}_\tau^{-1}) \leq \min\left\{\lambda_{\max}(\hat{Q}^{-1}), \tau_n^{-1}\right\} = O_p\left(\min\left\{n^{2\delta}, \tau_n^{-1}\right\}\right). \tag{5.17}$$

The degree of singularity of $\hat{Q}_\tau$ depends on the relative effect of weak instruments and penalization, namely, $\min\left\{n^{2\delta}, \tau_n^{-1}\right\}$. With (5.17), we can now derive the rate of convergence and asymptotic normality of $\hat{h}_\tau(\cdot)$.

# 6 Rate of Convergence

## 6.1 Assumptions

First we state regularity conditions under which we find the rate of convergence of the penalized series estimator introduced in the previous section. Recall that, throughout the paper, the orders of approximating functions for the structural function $g_0(\cdot)$, control function $\lambda_0(\cdot)$, and reduced-form function $\Pi_0(\cdot)$ depend on $n$ so that $K_1 = K_1(n)$, and $K_2 = K_2(n)$, and $L = L(n)$, respectively. And also $K = K_1 + K_2 - 1 = K(n)$. We consider the general case of $K_1 \neq K_2$, although we assume that $K_1$ and $K_2$ grow at the same rate, i.e., $K_1 \asymp K_2$, where $a_n \asymp b_n$ implies $a_n/b_n$ is bounded below and above by constants that are independent of $n$. Then we have $K \asymp K_1 \asymp K_2$. This setting can be justified by the assumption that the functions $g_0(\cdot)$ and $\lambda_0(\cdot)$ have the same smoothness, which is in fact imposed in Assumption C below. Also let $X = (x, z)$, and $f_u$ and $f_w$ be the density functions of $u = (z, v)$ and $w = (x, v)$, respectively.

**Assumption A (Random sample)** $\{(y_i, x_i, z_i) : i = 1, 2, ...\}$, *are i.i.d. and var*$(x|z)$ *and var*$(y|X)$ *are bounded functions of* $z$ *and* $X$*, respectively.*

As Newey (1997) points out, the bounded conditional variance assumption is difficult to relax without affecting the convergence rates.

**Assumption B** *$u = (z, v)$ is continuously distributed and the density of $u$ is bounded and bounded away from zero on $\mathcal{Z} \times \mathcal{V}$, where the support $\mathcal{Z}$ of $z$ is a Cartesian product of compact, connected intervals and the support $\mathcal{V}$ of $v$ is compact.*

The assumption is useful to bound below and above the eigenvalues of the transformed second moment matrix of approximating functions (i.e., $Q^*$). This condition is worth a discussion in the context of identification and weak instruments. An identification condition like Assumption ID2′ in Section 3.2 is embodied in this assumption. To see this, note that $f_u$ being bounded away from zero means that there is no functional relationship between $z$ and $v$, which in turn implies Assumption ID2′(a)(iii).[17] On the other hand, an assumption written in terms of $f_w$ like Assumption 2 in NPV (p.574) cannot be imposed here. Observe that $w = (\Pi(z) + v, v)$ depends on the behavior of $\Pi(\cdot)$, and hence $f_w$ is not bounded away from zero uniformly over $n$ under Assumption L; in fact, it approaches a singular density. Making use of the transformation matrix, we make an assumption in terms of $f_u$ and the effect of the weak instruments is handled separately.

The assumption for the Cartesian products of supports, namely $\mathcal{Z} \times \mathcal{V}$ and $\mathcal{X} \times \mathcal{V}$ (below), and the compactness of $\mathcal{V}$ in Assumption B can be replaced by introducing a trimming function as in NPV which ensures bounded rectangular supports. Lastly, similar to NPV, Assumption B can be weakened to hold only for some component of the distribution of $z$. And, one could allow some components of $z$ to be discrete, as long as they have finite supports.

Next, Assumption C is a smoothness assumption on the structural and reduced-form functions. Let $\mathcal{W} = \mathcal{X} \times \mathcal{V}$ be the support of $w = (x, v)$.

**Assumption C** *$g_0(x)$ and $\lambda_0(v)$ are in $L^2(F_x)$ and $L^2(F_v)$, respectively, and are Lipschitz and continuously differentiable of order $s$ on $\mathcal{W}$. $\Pi_0(z)$ is in $\mathcal{C}(\mathcal{Z})$ and is Lipschitz and continuously differentiable of order $s_1$ on $\mathcal{Z}$.*

This assumption ensures that the series approximation error shrinks as the number of approximating functions increases. It also motivates the introduction of penalization as well. Note that the same smoothness for $g_0(\cdot)$ and $\lambda_0(\cdot)$ is assumed because there is no systematic reason that one is particularly smoother than the other.

The next regularity condition restricts the rate of the growth of the number, $K$ and $L$, of approximating functions.

**Assumption D** *When the approximating functions are power series, $n^{2\delta}K^{7/2}[\sqrt{L/n}+L^{-s_1/d_z}] \to 0$, $n^{-\delta}K^{11/2} \to 0$, and $L^3/n \to 0$. When the approximating functions are regression splines, $n^{2\delta}K^2[\sqrt{L/n} + L^{-s_1/d_z}] \to 0$, $n^{-\delta}K^3 \to 0$, and $L^2/n \to 0$.*

This condition is stronger than the corresponding assumption in NPV (Assumption 4, p. 575) where weak instruments are not considered.

[17] For the definition of a functional relationship, see NPV (p.568).

### 6.2 Consistency and Rate of Convergence

First, we provide results for the rate of convergence in probability of the penalized series estimator $\hat{h}_\tau(w)$ for the true $h_0(w) = g_0(x) + \lambda_0(v)$, in terms of $L^2$ and uniform distance. Then, we give conditions for consistency. The convergence rate for $\hat{g}_\tau(x)$ is also derived. The rate of convergence of $\hat{h}_\tau(w)$ is of interest because the estimator of $\lambda_0(v)$ gives useful information about the model; see the Conclusion section for related discussion. Recall that $d_x$ is the dimension of $x$ and $d_z$ of $z$. Let $F_w(w) = F(w)$ be the distribution function of $w$.

**Theorem 6.1** *Suppose Assumptions A-D, and L are satisfied. Let* $R_n = \min\{n^\delta, \tau_n^{-1/2}\}$ *if* $\tau_n > 0$, *and* $R_n = n^\delta$ *if* $\tau_n = 0$. *Then,*

$$\left\{\int \left[\hat{h}_\tau(w) - h_0(w)\right]^2 dF(w)\right\}^{\frac{1}{2}} = O_p\left(R_n(\sqrt{K/n} + K^{-\frac{s}{d_x}} + \tau_n \cdot R_n + \sqrt{L/n} + L^{-\frac{s_1}{d_z}})\right).$$

*Also, with* $q = 1/2$ *for splines, and* $q = 1$ *for power series,*

$$\sup_{w\in\mathcal{W}}\left|\hat{h}_\tau(w) - h_0(w)\right| = O_p\left(R_n \cdot K^q(\sqrt{K/n} + K^{-\frac{s}{d_x}} + \tau_n \cdot R_n + \sqrt{L/n} + L^{-\frac{s_1}{d_z}})\right).$$

The proofs of the theorems are in the Appendix.

**Remarks:** 1. Suppose $\tau_n = 0$, which is the case with no penalization. Then, with $R_n = n^\delta$, Theorem 6.1 gives the rates of convergence of the unpenalized series estimator $\hat{h}(\cdot)$ defined in (5.4). For example, with $\|\cdot\|_{L^2}$ denoting the $L^2$ norm above,

$$\left\|\hat{h} - h_0\right\|_{L^2} = O_p\left(n^\delta(\sqrt{K/n} + K^{-s/d_x} + \sqrt{L/n} + L^{-s_1/d_z})\right).$$

Recall that $\delta$ measures the strength of instruments. When $\delta = 0$, i.e., in a strong instrument case, the rate coincides with that of NPV (p.575, Lemma 4.1).

2. Suppose $\tau_n = 0$ and $\delta \neq 0$. The rate deteriorates compared to the strong instrument case by $n^\delta$, the weak instrument rate. Note that the terms $\sqrt{K/n}$ and $K^{-s/d_x}$ correspond to the variance and bias of the second stage estimator, respectively, and $\sqrt{L/n}$ and $L^{-s_1/d_z}$ are those of the first stage estimator. The way that $n^\delta$ enters in the rates of Theorem 6.1 implies that the effect of weak instruments (hence multicollinearity) not only exacerbates the variance but also the bias. This is different from a parametric case where multicollinearity only results in imprecise estimates but does not introduce bias.

3. The symmetric effect of weak instruments on bias and variance featured in Remark 2 implies that the problem of weak instruments *cannot* be resolved by the choice of number of terms in the series estimator. This is also related to the discussion of Section 5.2 that the truncation method does not work as a regularization method for weak instruments.

4. More importantly, in the case where $\tau_n > 0$, the way that $R_n$ enters in the convergence rates implies that penalization can reduce both bias and variance by the same mechanism working in an opposite direction to the effect of weak instruments. Note that $\tau_n \cdot R_n$ is the penalty bias term, namely, the additional bias term introduced by penalization. It is, however,

not a serious issue since the additional bias can possibly be dominated by the existence of the series approximation bias terms, $K^{-s/d_x}$ and $L^{-s_1/d_z}$. Therefore, the penalization method used in nonparametric settings as in the present paper is fundamentally different from using ridge regression in linear model which produces a biased estimator (with smaller variance). In Monte Carlo simulations below, we simulate the finite sample performance of the penalized series estimator and show that variance as well as bias are reduced compared to the unpenalized estimator when instruments are weak.

5. When implementing the penalized series estimator in practice, there is a remaining issue of choosing tuning parameters, namely the penalization parameter $\tau_n$ and the order $K$ and $L$ of the series. In the simulations, we try out a few fixed values $\tau$ and choose the one that appears most reasonable. We also try different values of $K$ and $L$. A data-driven procedure such as the cross-validation method can also be used (Arlot and Celisse (2010)). This method is used for choosing $\tau$ in the empirical section below. There is, however, no optimal way of choosing the tuning parameters that is developed in the literature (Blundell, Chen and Kristensen (2007, pp. 1636-1637)). It is interesting to further investigate the sensitivity of the penalized estimator to the choice of $\tau$. Recall that $\hat{h}_\tau(\cdot) = p^K(\cdot)'\hat{\beta}_\tau = p^K(\cdot)'(\hat{Q} + \tau I)^{-1}\hat{P}'y/n$ is a function of $\tau$. To determine the sensitivity of $\hat{h}_\tau(\cdot)$ to $\tau$, we consider the sharp bound on $\lambda_{\max}((\hat{Q} + \tau I)^{-1})$ that appears in (5.11), namely, $[\lambda_{\min}(\hat{Q}) + \tau]^{-1}$. As a measure of the sensitivity, we calculate the absolute value of the first derivative of the bound:

$$\left| \frac{\partial[\lambda_{\min}(\hat{Q}) + \tau]^{-1}}{\partial \tau} \right| = \left[\lambda_{\min}(\hat{Q}) + \tau\right]^{-2}. \tag{6.1}$$

Note that the sensitivity is a decreasing function of $\lambda_{\min}(\hat{Q})$. That is, as the instruments become weaker (i.e., $\lambda_{\min}(\hat{Q})$ becomes smaller due to increased singularity), the performance of $\hat{h}_\tau(\cdot)$ becomes more sensitive to a change in $\tau$. This can have certain implications in practice.

6. As NPV point out, the dimension of the explanatory variables involved in the theorem is smaller than the dimension of $w$. This feature is by exploiting the additive structure of the model so that we can eliminate the interaction terms. Additional relevant discussion can be found in Assumption A.2 and p.472 of Andrews and Whang (1990), or Theorem 3.2 of Powell (1981, p. 26).

7. The nonparametric rate $\sqrt{L/n} + L^{-s_1/d_z}$ which appears in the convergence rates is due to generated regressors as the residuals $\hat{v}_i$ are obtained from the first stage nonparametric estimation.

8. Suppose $\tau_n > 0$. The convergence rate can be analyzed in two different cases according to the relative effect of weak instruments and penalization: (1) Weak instrument dominating case where $\min\{n^\delta, \tau_n^{-1/2}\} = n^\delta$: In this case, the $L^2$ convergence rate becomes

$$\left\| \hat{h}_\tau - h_0 \right\|_{L^2} = O_p\left(n^\delta(\sqrt{K/n} + K^{-s/d_x} + \sqrt{L/n} + L^{-s_1/d_z})\right) + O_p(n^{2\delta}\tau_n),$$

where $O_p(n^{2\delta}\tau_n) = o_p(1)$ since $n^\delta/\tau_n^{-1/2} = n^\delta\tau_n^{1/2} \to 0$ (considering only a strict case of the minimum function). Here, the rate is similar to the rate with unpenalized $\hat{h}$, and consistency can be achieved for a certain range of values of $\delta$. Intuitively, in $\hat{Q}_\tau = \hat{Q}+\tau_n I$, the penalization parameter is sufficiently "small" relative to nearly singular matrix $\hat{Q}$ that the resulting penalty bias term is $o_p(1)$; (2) Penalization dominating case where $\min\{n^\delta, \tau_n^{-1/2}\} = \tau_n^{-1/2}$: In this case the rate becomes

$$\left\| \hat{h}_\tau - h_0 \right\|_{L^2} = O_p\left( \tau_n^{-1/2}(\sqrt{K/n} + K^{-s/d_x} + \sqrt{L/n} + L^{-s_1/d_z}) \right) + O_p(1).$$

Here, the overall rate seems to be improved since the multiplying rate $\tau_n^{-1/2}$ is faster than the multiplying rate $n^\delta$ of the first case (or the rate with $\hat{h}$). Note that the penalty bias term is at worst of order $O_p(1)$. Without further restricting the model, however, we are not able to guarantee that the penalty bias shrinks to zero. It may be worth looking at the behavior of the penalty bias when more assumptions are imposed on the model or when different penalty functions other than the current $L^2$ penalty are considered. This topic, however, is beyond the scope of this paper. Case (2) is ruled out in the following analysis of consistency.

For a more concrete comparison between the rates $n^\delta$ and $\tau_n^{-1/2}$, let $\tau_n = n^{-2\delta_\tau}$, where $\delta_\tau > 0$. The larger $\delta_\tau$ is, the faster the penalization parameter converges to zero, and hence presumably the smaller is the additional bias. Then, (1) the weak instrument dominating case is when $\delta_\tau > \delta$, and (2) the penalization dominating case is when $0 < \delta_\tau < \delta$. As discussed above, in case (1), there is possible room for consistency, and hence it is necessary for consistency that $\delta$ satisfies $\delta < \delta_\tau$.

Next, we find the optimal $L^2$ convergence rate. With $K = n^{1/(1+2s/d_x)}$ and $L = n^{1/(1+2s_1/d_z)}$, the optimal convergence rate is $n^{-q}$ where $q = \min\left\{\frac{s}{1+2s/d_x}, \frac{s_1}{1+2s_1/d_z}\right\} - \delta$, since $R_n = n^\delta$. Note that, without weak instruments ($\delta = 0$), $n^{-s/(1+2s/d_x)}$ and $n^{-s_1/(1+2s_1/d_z)}$ are optimal rates of Stone (1982) for the second and first step estimation, respectively. With weak instruments ($\delta \neq 0$), optimal rates in the sense of Stone (1982) are not attainable.[18] Since $n^{-q} = o(1)$ implies consistency, condition $\delta < \min\left\{\frac{s}{1+2s/d_x}, \frac{s_1}{1+2s_1/d_z}\right\}$ is required for consistency.

The results are summarized and consistency is achieved in the following corollary to Theorem 6.1.

**Corollary 6.2 (Consistency)** *Suppose the Assumptions of Theorem 6.1 are satisfied. Let $K = O(n^{1/(1+2s/d_x)})$ and $L = O(n^{1/(1+2s_1/d_z)})$. If*

$$\delta < \min\left\{\delta_\tau, \frac{s}{1+2s/d_x}, \frac{s_1}{1+2s_1/d_z}\right\} \tag{6.2}$$

*then $\left\| \hat{h}_\tau - h_0 \right\|_{L^2} = o_p(1)$.*

9. For consistency of the penalized series estimator, the penalization parameter needs to shrink faster than the weak instrument effect, i.e., $\delta < \delta_\tau$. This result implies that when

[18] The uniform convergence rate does not attain Stone's (1982) bound even without the weak instrument factor (Newey (1997, p.151)).

instruments are weak (large $\delta$), there is less room to choose $\delta_\tau$ for consistency than the case when they are strong. This is related to the sensitivity issue of choosing the penalization parameter value, which is discussed above.

10. For consistency, if the structural function is "wiggly" (small $s$, hence small $\frac{s}{1+2s/d_x}$), the instrument should not be too weak at a given penalization parameter value. This is a trade-off between the smoothness of the structural function and the required strength of instruments. This, in turn, implies that the weak instrument problem can be mitigated with some smoothness restrictions, which is actually one of our justifications for introducing the penalization method. Conversely, if the true structural function of interest is less smooth, then stronger instruments are required for desirable performance of the estimator. Also, if the function is wiggly, there is more room to choose the penalization parameter value while consistency is guaranteed for given instruments.

11. In general, condition (6.2) implies that consistency is achieved when the instruments are only *mildly weak*.[19]

Theorem 6.1 leads to a subsequent theorem which focuses on the rate of convergence of the structural estimator $\hat{g}_\tau(\cdot)$ of $g_0(\cdot)$ after subtracting the constant term which is not identified.

**Theorem 6.3** *Suppose Assumptions A-D and L are satisfied. Let* $R_n = \min\{n^\delta, \tau_n^{-1/2}\}$ *if* $\tau_n > 0$, *and* $R_n = n^\delta$ *if* $\tau_n = 0$. *For* $\hat{\Delta}(x) = \hat{g}_\tau(x) - g_0(x)$,

$$\left\{\int \left[\hat{\Delta}(x) - \int \hat{\Delta}(x) dF(w)\right]^2 dF(w)\right\}^{\frac{1}{2}} = O_p\left(R_n\left[\sqrt{\frac{K}{n}} + K^{-\frac{s}{d_x}} + \tau_n \cdot R_n + \sqrt{\frac{L}{n}} + L^{-\frac{s_1}{d_z}}\right]\right).$$

*Also, if* $\hat{g}_\tau(x) = \hat{h}_\tau(x, \bar{v}) - \bar{\lambda}$ *and* $\bar{\lambda} = \lambda_0(\bar{v})$, *then, with* $q = 1/2$ *for splines, and* $q = 1$ *for power series,*

$$\sup_{x \in \mathcal{X}} |\hat{g}_\tau(x) - g_0(x)| = O_p\left(R_n K^q \left[\sqrt{K/n} + K^{-\frac{s}{d_x}} + \tau_n \cdot R_n + \sqrt{L/n} + L^{-\frac{s_1}{d_z}}\right]\right).$$

The consistency result for $\hat{g}_\tau(\cdot)$ can be derived analogously, which we omit here. The convergence rate is net of the constant term. We can further assume $E[\varepsilon] = 0$ to identify the constant.

The convergence rate results of this section (and also the asymptotic normality results below) can be applied in a special case of model (2.1), where the reduced-form equation is

[19] We can proceed further to derive an "effective" bound on $\delta$ for consistency. Suppose $\delta_\tau = 0$, $d_x = d_z = 1$. Note that Assumption D(a) implies $\delta < \frac{1}{4} - \frac{7}{4(1+2s)} - \frac{1}{4(1+2s_1)}$, since $n^{2\delta} K^{7/2}[L^{1/2} n^{-1/2} + L^{-s_1}] = n^{2\delta} n^{7/2(1+2s)} n^{1/2(1+2s_1)} n^{-1/2}$ by $K = O(n^{1/(1+2s)})$ and $L = O(n^{1/(1+2s_1)})$. Therefore the effective bound on $\delta$ is

$$\delta < \min\left\{\frac{s}{1+2s}, \frac{s_1}{1+2s_1}, \frac{1}{4}\left(1 - \frac{7}{1+2s} - \frac{1}{1+2s_1}\right)\right\}.$$

Let $s = s_1$ for simplicity, then we have consistency if $\delta < \min\left\{\frac{s}{1+2s}, \frac{1}{4}\left(1 - \frac{8}{1+2s}\right)\right\}$, which is the case of mildly weak instruments. Note that, when $s = 4$ (worst case), $\delta < 1/36$, and when $s \to \infty$ (best case), $\frac{s}{1+2s} \to 1/2$ so $\delta < 1/4$.

linear with vector $z$ where its coefficients have different (or the same) drifting rates. In this case, asymptotic results can be obtained in a straightforward fashion. Consider a penalized series estimator $\hat{h}_\tau(\cdot)$ that is defined as in Section 5.3, but with linear least squares residuals. Then, regarding the convergence rate of the estimator, the nonparametric rate of the first stage ($\sqrt{L/n}+L^{-s_1/d_z}$) disappears from the rates in Theorem 6.1 or 6.3, since the parametric rate ($1/\sqrt{n}$) is dominated by the second stage nonparametric rate. The weak instrument and penalization components ($R_n$) still remain.

### 6.3 Linear versus Nonparametric Reduced Form

Here, we discuss one of the practical implications of the identification and asymptotic results of this paper, which concerns the specification of the reduced form in nonparametric triangular models. In applied research that uses nonparametric triangular models, a linear specification of the reduced form is largely prevalent; see the Introduction. The linear specification might be introduced either to avoid the curse of dimensionality with many covariates at hand, because the nonparametric structural equation is of primary interest, or simply, for the ad hoc reason that it is easy to implement. A linear specification of the reduced form, however, may be less desirable, if not more harmful, than is generally expected when one nonparametrically estimates the structural function in the outcome equation.

The rank condition derived in the identification analysis, i.e., $\Pr[\partial\Pi_0(z)/\partial z \neq 0] > 0$ in the univariate $x$ and $z$ case, suggests that a small region where $x$ and $z$ are relevant contributes to the identification of $g(\cdot,\cdot)$. This implies that identification power can be enhanced by exploiting the entire nonlinear relationship between $x$ and $z$. When the reduced form is linearly specified, any true nonlinear relationship is "flattened out" and the situation is more likely to have the problem of weak instruments.[TECHNICAL DETAILS WILL BE ADDED HERE.]

As an illustration, consider a situation where a nonparametric reduced form results in strong identification while a linear specification produces weak instruments. By Theorem 6.1 with an unpenalized estimator ($\tau_n = 0$) and univariate $x$ and $z$ for ease of exposition, with the nonparametric reduced form, the convergence rate is

$$\left\|\hat{h} - h_0\right\|_{L^2} = O_p\left(\sqrt{K/n} + K^{-s} + \sqrt{L/n} + L^{-s_1}\right). \tag{6.3}$$

On the other hand, with the linear reduced form, the rate is (even with pseudo-true $h(\cdot)$)

$$\left\|\hat{h} - h_{pseudo}\right\|_{L^2} = O_p\left(n^{\delta}[\sqrt{K/n} + K^{-s}]\right), \tag{6.4}$$

where the first-stage rate disappears as previously discussed.

Note that in (6.3), the rate $\sqrt{L/n} + L^{-s_1}$ from the nonparametric first stage estimation is not likely to worsen the rate since the rate $\sqrt{K/n} + K^{-s}$ from the nonparametric second stage is already present. This argument becomes even more obvious when we assume $s = s_1$ (i.e., the reduced form $\Pi_0(\cdot)$ and outcome function $h_0(\cdot)$ are equally smooth) and $K = L$ (i.e., we use the same number of approximating functions to estimate them), and those two rates coincide.

In sum, a nonparametric specification of the reduced form has advantages. When the true association of $x$ and $z$ is nonlinear, a linear specification is subject to weak instruments more than the nonparametric specification is, let alone the problem of misspecification. Hence it is more likely to exacerbate the bias and variance performance of the resulting estimators. On the other hand, one can achieve a significant gain in the performance by nonparametrically estimating the relationship between $x$ and $z$ without significant loss. Meanwhile, when the true relationship of $x$ and $z$ is linear, which often seems implausible in most empirical examples, both nonparametric and linear specifications of the reduced form will result in a similar rate. Therefore, if there are no economic or any theoretical reasonings for specifying the reduced form in one way or another, we recommend having a nonparametric reduced form. Note that the dimensionality problem can be dealt with by a single-index model or semiparametric model of the reduced form. If one has a theory or economic justification about the true relationship between $x$ and $z$ and if it is linear, then one needs to be more cautious about weak instruments than when the truth is believed to be nonlinear.

A nonparametric reduced form exploits the nonlinear relationship between $x$ and $z$, and hence enhances identification power. This phenomenon might be interpreted in terms of the "optimal instruments" in GMM settings of Amemiya (1977); see also Newey (1990) and Jun and Pinkse (2007). For a rigorous analysis, however, we may need different frameworks (cf. Newey and Powell (2003), Ai and Chen (2003)), and this topic is beyond the scope of this paper.

# 7 Asymptotic Distributions

In this section we establish the asymptotic normality of linear functionals of the penalized series estimator $\hat{h}_\tau(\cdot)$. We consider two types of linear functionals of $h_0(\cdot)$, namely, $h_0(\cdot)$ at a certain value (i.e., $h_0(\bar{w})$) and the weighted average derivative of $h_0(\cdot)$ (i.e., $\int v(w)[\partial h_0(w)/\partial x]dw$). The linear functionals of $h(\cdot)$ are denoted as $a(h)$. Then, the estimator $\hat{\theta}_\tau = a(\hat{h}_\tau)$ of $\theta_0 = a(h_0)$ is the natural "plug-in" estimator. Let $A = (a(p_{1K}), a(p_{2K}), .., a(p_{KK}))$, where $p_{jK}(\cdot)$ is an element of $p^K(\cdot)$. Then

$$\hat{\theta}_\tau = a(\hat{h}_\tau) = a(p^K(x)'\hat{\beta}_\tau) = A\hat{\beta}_\tau,$$

which implies that $\hat{\theta}_\tau$ is a linear combination of the two-step least squares estimators $\hat{\beta}_\tau$. Then, the following variance estimator of $a(\hat{h}_\tau)$ can be naturally defined:

$$\begin{aligned}
\hat{V}_\tau &= A\hat{Q}_\tau^{-1}\left(\hat{\Sigma}_\tau + \hat{H}_\tau \hat{Q}_1^{-1}\hat{\Sigma}_1\hat{Q}_1^{-1}\hat{H}_\tau'\right)\hat{Q}_\tau^{-1}A', \\
\hat{\Sigma}_\tau &= \sum_{i=1}^n p^K(\hat{w}_i)p^K(\hat{w}_i)'[y_i - \hat{h}_\tau(\hat{w}_i)]^2/n, \qquad \hat{\Sigma}_1 = \sum_{i=1}^n \hat{v}_i^2 r^L(z_i)r^L(z_i)'/n, \\
\hat{H}_\tau &= \sum_{i=1}^n p^K(\hat{w}_i)\left\{\left[\partial\hat{h}_\tau(\hat{w}_i)/\partial w\right]'\partial w(X_i,\hat{\Pi}(z_i))/\partial\pi\right\}r^L(z_i)'/n, \qquad \hat{Q}_1 = R'R/n,
\end{aligned}$$

where $X$ is a vector of variables that includes $x$ and $z$, and $w(X,\pi)$ is a vector of functions of $X$ and $\pi$, where $\pi$ is a possible value of $\Pi(z)$.

The following are additional regularity conditions for the asymptotic normality of $a(\hat{h}_\tau)$.

**Assumption E** *$\sigma^2(X) = var(y|X)$ is bounded away from zero, $E[\eta^4|X]$ is bounded, and $E[v^4|X]$ is bounded. Also, $h_0(w)$ is twice continuously differentiable in $v$ with bounded first and second derivatives, where $w = (x,v)$.*

This assumption strengthens the boundedness of conditional second moments in Assumption A. For the next assumption, let $d_w = 2d_x$, $|\mu| = \sum_{j=1}^{d_w}\mu_j$ for a $d_w$-vector $\mu$, and, for a nonnegative integer $r$, let $|h|_r = \max_{|\mu|\le r}\sup_{w\in\mathcal{W}}|\partial^\mu h(w)|$, where $\partial^\mu h(w) = \partial^{|\mu|}h(w)/(\partial w_1^{\mu_1}\partial w_2^{\mu_2}\cdots\partial w_{d_w}^{\mu_{d_w}})$.

**Assumption F** *Either (a) or (b) holds. (a) There exists $\nu(w)$ and $\beta_K$ such that $E[\|\nu(w)\|^2] < \infty$, $a(h_0) = E[\nu(w)h_0(w)]$, $a(p_j) = E[\nu(w)p_j(w)]$, $E[\|\nu(w) - p^K(w)'\beta_K\|^2] \to 0$ and $K \to \infty$; (b) $a(h)$ is a scalar, $|a(h)| \le |h|_r$ for some $r \ge 0$, and there exists $\beta_K$ such that as $K \to \infty$, $a(p^{K\prime}\beta_K)$ is bounded away from zero while $E[\{p^K(w)'\beta_K\}^2] \to 0$.*

Note that Assumption F(a) includes the case of the weighted average derivative of $h(\cdot)$, and F(b) the case of $h(\cdot)$ at a certain value. Under Assumption F(a), $\sqrt{n}$-consistency is achieved for the estimator $\hat{\theta}$. Let $\rho(z) = E[\nu(w)\partial h_0(w)/\partial v'|z]$. Then, the asymptotic variance of $\hat{\theta}$ in this case can be expressed as

$$\bar{V} = E\left[\nu(w)\nu(w)'var(y|X)\right] + E\left[\rho(z)var(x|z)\rho(z)'\right].$$

The next condition restricts the rate of growth of $K$ and $L$.

**Assumption G** *As $n \to \infty$, it holds that $\sqrt{n}K^{-s/d_x} \to 0$, $\sqrt{n}L^{-s_1/d_z} \to 0$, and $n^{\delta+\frac{1}{2}}\tau_n \to 0$; also, for power series, $n^{3(\delta-\frac{1}{6})}K^4L^{1/2} \to 0$, $n^{\delta-\frac{1}{2}}\{K^3L^{3/2} + K^4L^{1/2}(K+L)^{1/2}\} \to 0$, and $n^{4(\delta-\frac{1}{4})}\{K^3+L^3\} \to 0$, and, for splines, $n^{3(\delta-\frac{1}{6})}K^{5/2}L^{1/2} \to 0$, $n^{\delta-\frac{1}{2}}\{K^{3/2}L^{3/2} + K^2L^{1/2}(K+L)^{1/2}\} \to 0$, and $n^{4(\delta-\frac{1}{4})}\{K^2+L^2\} \to 0$.*

The first two conditions $\sqrt{n}K^{-s/d_x} \to 0$ and $\sqrt{n}L^{-s_1/d_z} \to 0$ in Assumption G implies "overfitting" in that the bias ($K^{-\alpha}$) shrinks faster than $1/\sqrt{n}$, the usual rate of standard deviation of the estimator. As in NPV, we introduce overfitting for inference. This assumption is stronger than that in NPV (Assumption 8, p. 582). The value of $\delta$ which satisfies G needs to be small, implying that the instruments are at most "mildly" weak to obtain the asymptotic normality of $\hat{\theta}_{\tau}$. Also, $\tau_n$ needs to converge fast enough to satisfy $n^{\delta+\frac{1}{2}}\tau_n \to 0$ in G. Given these assumptions, we establish the asymptotic normality for the functionals of the penalized series estimator. Let $\zeta_r^v(K) = \max_{|\mu|\leq r} \sup_{v\in\mathcal{V}} \left\|\partial^{\mu} p^K(v)\right\|$.

**Theorem 7.1** *If Assumptions A-G and L are satisfied, then* $\hat{\theta}_{\tau} = \theta_0 + O_p(n^{2(\delta-\frac{1}{4})}\zeta_r^v(K))$ *and*

$$\sqrt{n}\hat{V}_{\tau}^{-1/2}(\hat{\theta}_{\tau} - \theta_0) \to_d N(0,1).$$

*Furthermore, if Assumption F(a) is satisfied,*

$$\sqrt{n}(\hat{\theta}_{\tau} - \theta_0) \to_d N(0,\bar{V})$$

*and* $\hat{V}_{\tau} \to_p \bar{V}$.

Besides asymptotic normality, the results also provide the bound on the convergence rate of $\hat{\theta}_{\tau}$ as well as $\sqrt{n}$-consistency. Note that $\sqrt{n}$-consistency is achieved for the weighted average derivative estimator. The results of this theorem are similar to those of NPV, except that the bound on the rate achieved here is slower. The fact that the rate is slower (i.e., the multiplier of $(\hat{\theta}_{\tau} - \theta_0)$ is of smaller order) in the case of weak instruments can be seen as if the effective sample size is small. One can give similar discussions as in the convergence rate part that there are certain ranges of the strength of instruments and value of the penalization parameter that guarantee the asymptotic normality of $\hat{\theta}_{\tau}$.

There still remain issues when the results of Theorem 7.1 are used for inference, e.g., by constructing pointwise asymptotic confidence intervals. As long as nuisance parameters are present, such an inferential procedure may depend on the strength of instruments or on the choice of penalization parameter. Robust inference against weak instruments in nonparametric models can be an interesting future research topic.

Lastly, one closely related paper to the asymptotic results of the present paper is Jiang, Fan and Fan (2010). In an additive nonparametric regression model where the regressors are highly correlated, they establish pointwise asymptotic normality for local linear and integral estimators. Their results show the way that the correlated regressors affect bias and variance. Once our problem is transformed to an additive nonparametric regression with multicollinearity (Section 5.2), we can develop weak instrument asymptotic theory for local linear and integral estimators of functionals of $h_0(\cdot)$. The results of Jiang, Fan and Fan (2010), however, cannot be automatically applied since generated regressors ($\hat{v}_i$) are involved in our problem.

## 8 Monte Carlo Simulations

Before we apply the estimation strategies developed in this paper to a real world example, in this simulation section, we investigate the problems of weak instruments in nonparametric estimation and demonstrate the finite sample performance of the penalized estimator.

We are particularly interested in the finite sample distribution and bias and variance of the usual IV estimator obtained by the control function approach, that is, the two step estimation procedure of Section 5.1. For bias and variance performance, at a wide range of strength of instruments, we compare the IV estimators with least squares (LS) estimators which ignore the endogeneity. We are also interested in the finite sample gain in terms of bias and variance of IV estimators obtained by penalization ("penalized IV estimators") discussed in Section 5.3 compared to the unpenalized estimators.

The simulation results can be summarized as follows. Even with a strong instrument in a conventional sense, the finite sample distribution of the unpenalized IV estimator is far from normal. Also in this case, unpenalized IV estimators do poorly in terms of mean squared errors (MSE) compared to LS estimators. Variance seems to be the bigger problem but bias is also worrisome. Penalization alleviates much of the variance problem induced by the weak instrument, and also surprisingly works well on bias for a range of the weak instrument. Even when the instrument is strong, the bias of the penalized IV estimator is no larger than that of the unpenalized one.

## 8.1 Simulation Design

We consider the following data generation process (DGP): With the true functional form of $g(x)$ being $g_0(x) = \Phi(\frac{x-\mu_x}{\sigma_x})$,

$$y = \Phi(\frac{x-\mu_x}{\sigma_x}) + \varepsilon, \qquad x = \pi_1 + z\pi + v,$$

where $y$, $x$, and $z$ are univariate, $z \sim N(\mu_z, \sigma_z^2)$ with $\mu_z = 0$ and $\sigma_z^2 = 1$, and $(\varepsilon, v)' \sim N(0, \Sigma)$ with $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. Note that $|\rho|$ measures the degree of endogeneity, and we consider $\rho \in \{0.2, 0.5, 0.95\}$. Due to the bivariate normal assumption for $(\varepsilon, v)'$, we are implicitly imposing linearity in the function $E[\varepsilon|v] = \lambda(v)$. The sample $\{z_i, \varepsilon_i, v_i\}$ is i.i.d. with size $n = 1000$. The number of simulation repetitions is $s \in \{500, 1000, 5000\}$.

We consider different strengths of the instrument by considering different values of $\pi$. Let the intercept $\pi_1 = \mu_x - \pi\mu_z$ with $\mu_x = 2$, so that $E[x] = \mu_x$ does not depend on the different choice of $\pi$. Note that $\sigma_x^2 = \pi^2\sigma_z^2 + 1$ still depends on $\pi$, which is reasonable as the signal contributed to the total variation of $x$ is a function of $\pi$. More specifically, to measure the strength of the instrument, we define the concentration parameter (Stock and Yogo (2005)):

$$\mu^2 = \frac{\pi^2 \sum_{i=1}^n z_i^2}{\sigma_v^2}$$

Note that, since the dimension of $z$ is one, the concentration parameter value and the first stage $F$-statistic are similar to each other.[20] The candidate values of $\mu^2$ are $\{4, 8, 16, 32, 64, 128, 256\}$,

[20] For example, in Staiger and Stock (1997), for $F = 30.53$ (strong instrument) a 97.5% confidence interval for $\mu^2$ is $[17.3, 45.8]$ and for $F = 4.747$ (weak instrument) a confidence interval for $\mu^2$ is $[2.26, 5.64]$.

which range from a weak to strong instrument in the conventional sense. As for the penalization parameter $\tau$, we consider candidate values of $\{0.001, 0.005, 0.01, 0.05., 0.1\}$.

The approximating functions used for $g_0(x)$ and $\lambda_0(v)$ are polynomials and p-spline with different choices of $(K_1, K_2)$, where $K_1$ is the number of terms for $g_0(\cdot)$ and $K_2$ for $\lambda_0(\cdot)$. Since $g_0(\cdot)$ and $\lambda_0(\cdot)$ are separately identified only up to an additive constant, we introduce normalization $\lambda_0(1) = \rho$, where $\rho$ is chosen because of the joint normality of $(\varepsilon, v)$. Then, $g_0(x) = h_0(x, 1) - \rho$.

## 8.2 Simulation Results

In the first part of the simulation, we calculate $\hat{g}_\tau(\cdot)$ and $\hat{g}_0(\cdot)$, the penalized series estimates and unpenalized series estimates, respectively, and compare their performances. For different strengths of the instrument, we compute estimates with different values of the penalization parameter. We choose $K_1 = K_2 = 6$ for polynomials as approximating functions, and $\rho = 0.5$. As one might expect, the choice of orders of the series is not significant as long as we are only interested in comparing $\hat{g}_\tau(\cdot)$ and $\hat{g}(\cdot)$.

Figures 4-7 present some representative results. Other results are similar and hence are omitted to save space. In Figure 4 we plot mean of $\hat{g}_\tau(\cdot)$ and $\hat{g}(\cdot)$ with concentration parameter $\mu^2 = 16$ and penalization parameter $\tau = 0.001$. The (blue) dotted-dash line is the true $g_0(\cdot)$. The (black) solid line is the (simulated) mean of $\hat{g}(\cdot)$ with the dotted band representing the 0.025-0.975 quantile ranges. Note that the difference between $g_0(\cdot)$ and the mean of $\hat{g}(\cdot)$ is the (simulated) bias. The (red) solid line is the mean of $\hat{g}_\tau(\cdot)$ with the dashed 0.025-0.975 quantile ranges. The plots for the unpenalized estimator indicate that with the given strength of the instrument, the variance is very large, which implies that functions with any trends can fit within the band; it indicates that the bias is also large.

The plots for the penalized estimator imply that the penalization significantly reduces the variance so that at least the upward trend of the true $g_0(\cdot)$ is now ensured. Note that the penalization corrects the bias in this case. Although $\mu^2 = 16$ is considered to be strong according to the conventional criterion, this range of the concentration parameter value can be seen as the case where the instrument is "nonparametrically" weak in the sense that the penalization induces a significant difference between $\hat{g}_\tau(\cdot)$ and $\hat{g}(\cdot)$. Figure 5 is with $\mu^2 = 256$, while other things are the same. In this case, the penalization induces no significant difference between $\hat{g}_\tau(\cdot)$ and $\hat{g}(\cdot)$. This can be seen as the case where the instrument is "nonparametrically" strong. It is noteworthy that the bias of the penalized estimator is no larger than the unpenalized one even in this case.

Figures 6-7 present the same plots but with penalization parameter $\tau = 0.005$. The patterns are similar to those of the previous case. Also, the comparison between Figures 4-5 and Figures 6-7 shows that the results are more sensitive to the change of $\tau$ in the weak instrument case than in the strong instrument case. This provides evidence for the theoretical discussion on sensitivity; see (6.1) in Section 6.2.

The fact that the penalized and unpenalized estimates significantly differ when the instrument is weak has a practical implication. That is, practitioners can be informed about whether the instrument they are using is worryingly weak by comparing penalized series estimates with unpenalized estimates. Similar approach can be found in linear weak instruments literature; for example, the biased TSLS estimates and the approximately median-unbiased LIML estimates of Staiger and Stock (1997) can be compared to detect weak instruments.

In Figures 8-9, we calculate a functional of the unpenalized series estimates, namely $\hat{\theta} = a(\hat{g}) = \hat{g}(\mu_x)$. With polynomials as the approximating functions, we use $K_1 = K_2 = 5$, and $\rho = 0.5$. Figures 8-9 compare the simulated distributions of $\hat{\theta}$ with two normal distributions, centered at $E\hat{\theta}$ ("normal_1") and $\theta_0 = a(g_0)$ ("normal_2"), respectively, where the difference of the two normal distributions thus indicates bias. When the instrument is nonparametrically weak (Figure 8), the shape of the finite sample distribution of $\hat{\theta}$ is far from its asymptotic normal distribution (normal_2), which implies that standard inference based on the normal critical values will result in size distortions. Also note that bias is present here. When the instrument is nonparametrically strong (Figure 9), the shape almost coincides with that of a normal distribution and bias is negligible.

Tables 2 reports the integrated squared bias, integrated variance and integrated MSE (Blundell, Chen and Kristensen (2007, p. 1638)) of penalized and unpenalized IV estimators and LS estimators of $g_0(\cdot)$. The LS estimates are calculated by series estimation of the outcome equation (with orders $K_1$) ignoring the endogeneity. We also calculate relative integrated squared bias (Staiger and Stock (1997)) and relative integrated MSE for better comparisons. Here, we use $K_1 = 6$ and $K_2 = 3$ in order to better reflect the fact that $\lambda_0(\cdot)$ is smoother than $g_0(\cdot)$ due to the joint normality of $(\varepsilon, v)$. The approximating functions are polynomials and $\rho = 0.5$. Results with different choices of orders $K_1$ and $K_2$ and degree of endogeneity $\rho$ show similar patterns; results with splines as approximating functions also show similar patterns; hence they are omitted here. Note that the usual bias and variance trade-offs are present as the order of the series changes.

The first three rows of entries in Table 2 are for the unpenalized IV estimator $\hat{g}(\cdot)$. As the instrument becomes weaker, the bias and variance of the IV estimator increase with greater proportion in variance. Next, the ratios of integrated bias and integrated MSE between the IV and LS estimators ($Bias^2_{IV}/Bias^2_{LS}$ and $MSE_{IV}/MSE_{LS}$ in Table 2) indicate the relative performance of the IV estimator compared to LS estimator. A ratio *bigger than unity* implies that the IV estimator performs *worse* than LS. In the table, the IV estimator does poorly in terms of MSE even when $\mu^2 = 16$, which is in the range of conventionally strong instruments; therefore, this can be considered as the case where the instrument is nonparametrically weak.

The following three rows in Table 2 are results for the penalized IV (PIV) estimator $\hat{g}_{\tau}(\cdot)$. Overall, the bias and variance are reduced and the decrease is significant for the variance. Obviously, the penalized IV estimator performs better than the LS estimator. More importantly, the last two rows ($Bias^2_{PIV}/Bias^2_{IV}$ and $MSE^2_{PIV}/MSE^2_{IV}$) suggest that, overall, the penalized IV estimator outperforms the unpenalized IV estimator in terms of bias and MSE. For example, when $\mu^2 = 8$, the bias of the penalized estimator is only about 0.4%

of the bias of the unpenalized ones, and the MSE of the penalized estimator is only about 1.4% of the MSE of the unpenalized one. Note that even in the situation of a fairly strong instrument (e.g., $\mu^2 = 256$), the bias of the penalized estimator does not exceed the bias of the unpenalized one. This provides evidence for the theoretical discussion that the penalty bias can be dominated by the existing series estimation bias; see Remark 4 in Section 6.2.

# 9 Application: Effect of Class Size

To illustrate our approach and apply the theoretical findings, we nonparametrically estimate the effect of class size on students' test scores. Among school inputs that affect students' performance, class size is thought to be easier to manipulate, and estimating its effect has been an interesting topic in the schooling literature. Angrist and Lavy (1999) analyze the effect of class size on students' reading and math scores in Israeli primary schools. With linear models, they find that the estimated effect is negative in most of the specifications they consider.

Here, we want to see whether the results of Angrist and Lavy (1999) are driven by their parametric assumptions. It is reasonable to consider a nonlinear effect of class size, since it is unlikely that the marginal effect is constant across class-size levels. Therefore, we consider nonparametrically extending their linear model: for school $s$ and class $c$,

$$score_{sc} = g(classize_{sc}, disadv_{sc}) + \alpha_s + \varepsilon_{sc}.$$

where $score_{sc}$ is the average test score within class, $classize_{sc}$ is the class size, $disadv_{sc}$ is the fraction of disadvantaged students, and $\alpha_s$ is an unobserved school-specific effect. Note that this model allows for different patterns for different subgroups of school/class characteristics (here, $disadv_{sc}$).

Class size is endogenous because it results from choices made by parents, schooling providers or legislatures, and hence is correlated with other determinants of student achievement. Angrist and Lavy (1999) use Maimonides' rule on maximum class size in Israeli schools to construct an IV. According to the rule, class size increases one-for-one with enrollment until 40 students are enrolled, but when 41 students are enrolled, the class size is dropped to an average of 20.5 students. Similarly, classes are split when the enrollment reaches 80, 120, 160, and so on, so that each size does not exceed 40. This rule can be expressed by the following nonlinear function of enrollment, which produces the IV (denoted as $f_{sc}$ following their notation):

$$f_{sc} = \frac{e_s}{int((e_s - 1)/40) + 1}$$

where $e_s$ is beginning-of-the-year enrollment count. This rule generates discontinuity in the enrollment / class-size relationship, which serves as exogenous variation. Figure 10 (Figure 1 in Angrist and Lavy (1999, p. 541)) depicts this relationship, where the class size induced by Maimonides' rule ($f_{sc}$) and the actual class size ($classize_{sc}$) are plotted by initial enrollment count ($e_s$). Note that with the sample plus and minus 7 students around the discontinuity points, IV exogeneity is more credible in addressing the endogeneity issue. Angrist and Lavy (1999) consider a linear reduced form:

$$classize_{sc} = \pi_0 + \pi_1 \cdot f_{sc} + \pi_2 \cdot disadv_{sc} + \tilde{v}_{sc}.$$

The dataset we use is from Angrist and Lavy (1999), which is the 1991 Israel Central Bureau of Statistics survey of Israeli public schools. We only consider the 4th graders. The sample size is $n = 2019$ for the full sample and 650 for the discontinuity sample. Given a linear reduced form, first stage tests have $F = 191.66$ with the discontinuity sample ($\pm 7$ students around the discontinuity points) and $F = 2150.4$ with the full sample. Lessons from the theoretical analyses above suggest that a strong instrument ($F = 191.66$) in a conventional sense can be weak in nonparametric estimation of the class-size effect, and a nonparametric reduced form can enhance the identification power. Therefore, we consider a nonparametric reduced form,

$$classize_{sc} = \Pi(f_{sc}, disadv_{sc}) + v_{sc}.$$

The sample is clustered, an aspect which is reflected in $\alpha_s$ of the outcome equation. Hence we use the block bootstrap when computing standard errors and take schools as bootstrap sampling units to preserve within-cluster (school) correlation. This produces cluster-robust standard errors. We use $b = 500$ bootstrap repetitions.

With the same example and dataset (full sample), Horowitz (2011, Section 5.2) uses the model and assumptions of the NPIV approach to nonparametrically estimate the effect of class size. To solve the ill–posed inverse problem, he conducts regularization by replacing the operator with a finite-dimensional approximation. First, we compare the NPIV estimate of Horowitz (2011) with the IV estimate obtained by the control function approach of this paper. Figure 11 (Figure 3 in Horowitz (2011, p. 375)) is the NPIV estimate of the function of class size ($g(\cdot,\cdot)$) for $disadv = 1.5(\%)$ with the full sample. The solid line is the estimate of $g$ and the dots show the cluster-robust 95% confidence band. As is noted in his paper, "the result suggests that the data and the instrumental variable assumption, by themselves, are uninformative about the form of any dependence of test scores on class size."

Figure 12 is the (unpenalized) IV estimate calculated using the triangular model (2.1) and the control function approach. Note that to facilitate the comparison with the NPIV estimate, we consider a nonparametric reduced form. Since NPIV approach does not specify any reduced-form relationship, it is reasonable to consider a flexible reduced form in the control function approach. The sample, the orders of the series and the value of *disadv* are the same as those for the NPIV estimate. The dashed lines in the figure are also the cluster-robust 95% confidence band. The result clearly presents a nonlinear shape of the effect of class size and suggests that the marginal effect diminishes as class size increases. Also the overall trend seems to be negative, which is consistent with the results of Angrist and Lavy (1999). The control function and NPIV approaches maintain different sets of assumptions (e.g., different orthogonality conditions for the IV), hence this comparison does not imply that one estimate performs better than the other. It does, however, imply that if the control function assumptions are reasonable, then they lead the data to be informative about the relationship of interest. Note that the assumption $E[\varepsilon_{sc}|v_{sc}, f_{sc}] = E[\varepsilon_{sc}|v_{sc}]$ is satisfied if $(\varepsilon_{sc}, v_{sc})$ are jointly independent of $f_{sc}$. Here, $\varepsilon_{sc}$ captures factors that determine the average test scores of a class other than the class size and the fraction of disadvantaged students, and $v_{sc}$ captures other factors that are correlated with enrolment.

Also, since the NPIV approach suffers from the ill-posed inverse problem even without the problem of weak instruments, the control function approach may be a more appealing framework than the NPIV approach in the presence of weak instruments.

We proceed by calculating penalized IV estimates by the estimation method of this paper. For all cases below, we find estimates for $disadv = 1.5(\%)$ as before. We use the discontinuity sample, where the instrument is possibly weak in this nonparametric setting. For comparison, however, we also calculate penalized IV estimates with the full sample, which has a much stronger instrument. We randomly select a subsample of the full sample to match the sample size with the discontinuity sample ($n = 650$). For the penalization parameter $\tau$, we use cross-validation, which is a data-driven procedure, to choose values among $\{0.005, 0.01, 0.015, 0.02, 0.05\}$.[21] The following results of cross-validation (Table 3) suggest that $\tau = 0.015$ is the MSE-minimizing value. We penalize $\beta_{-} = (\beta_2, \beta_3, ..., \beta_K)$ to effectively incorporate the smoothness.

Figure 13 depicts the penalized and unpenalized IV estimates with the discontinuity sample. There is a certain difference in the estimates, but the amount is small. It is possible that either the instrument is not very weak in this example or that cross-validation chooses a smaller value of $\tau$.

Lastly, in Figures 14-15 we plot IV estimates calculated using nonparametric and linear reduced forms with different choices for the number of terms in the series in different subfigures. Note that $K_1$, $K_2$, and $K_3$ are the number of terms for $g(\cdot)$, $\lambda(\cdot)$, and $\Pi(\cdot)$, respectively. Penalization is not considered in this analysis. Figure 14 depicts the case where the instrument is "nonparametrically" weak (with the discontinuity sample), and Figure 15 depicts where it is strong (with the full sample). With the weak instrument, the estimates with nonparametric reduced form are notably different from those with linear reduced form, presenting a flatter trend. This feature is true across different choices of the orders. On the other hand, with the strong instrument, there is no notable difference between the two. These results may indicate that the flexible reduced form exploits the nonlinear relationship of the instrument and the class size, strengthens the instrument, and hence results in more accurate estimates. When the instrument is strong, this exploitation does not result in a significant difference because a strong relationship is already captured by the linear specification.

[21] Specifically, we use 10-fold CV. See Arlot and Celisse (2010) for details.

# 10 Conclusions

This paper analyzes identification, estimation, and inference in a nonparametric triangular model in the presence of weak instruments. With a mild support condition, we derive a necessary and sufficient rank condition for identification, based on which we define the concept of nonparametric weak instruments. In estimation, we relate the weak instrument problem with the ill-posed inverse problem in the literature, and introduce penalization as a regularization scheme to alleviate the effects of weak instruments. We derive local asymptotics of the resulting penalized series estimator for the full range of strengths of instruments. The paper also provides lessons for applied researchers: IV can do "more harm than good" (Bound, Jaeger and Becker (1995)) to a greater extent in nonparametric models than it does in linear models, and hence further attention needs to be paid to the relevance condition of IV in nonparametric settings; penalized estimators can help solve the weak instrument problem; and nonparametric specification of the reduced form can be helpful when the structural function is nonparametric.

The findings and implications of this paper are not restricted to the present additively separable triangular models. The results can be adapted to the nonparametric limited-dependent-variable framework of Das, Newey and Vella (2003) and Blundell and Powell (2004). Weak instruments can also be studied in other nonparametric models with endogeneity, such as the IV quantile regression model of Chernozhukov and Hansen (2005) and Lee (2007).

Also, the results of this paper are directly applicable in several semiparametric specifications of the model of the present paper. With a large number of covariates, one can consider a semiparametric outcome equation or reduced form that is additively nonparametric in some components and parametric in others. One can also consider a single-index model for one equation or the other. As more structure is imposed on the model, the identification condition of Section 3.2 and the regularity condition of Section 6.1 can be weakened. Note that when the reduced form is of a single-index structure as $\Pi(z'\gamma)$, the strength of instruments is determined by the combination of $\partial\Pi(\cdot)/\partial(z'\gamma)$ and $\gamma$. Another example is where the structural function is parametric, while the control function remains nonparametric. A nonparametric reduced form is still appealing in that case by similar arguments to those made in Section 6.3.

Other subsequent research can be done concerning two specification tests: a test for relevance of the instruments and a test for endogeneity. These tests can be conducted by adapting the existing literature on specification tests where the test statistics can be constructed using the series estimators of this paper; see, e.g., Hong and White (1995). Testing whether instruments are relevant can be conducted with the nonparametric reduced-form estimate $\hat{\Pi}(\cdot)$. A possible null hypothesis is $H_0 : \Pr\{\Pi(z) = \mathit{const.}\} = 1$, which is motivated by our rank condition for identification. Testing whether the model suffers endogeneity problems can be conducted with the control function estimate $\hat{\lambda}(\cdot)$ obtained from $\hat{h}(w) = \hat{g}(x) + \hat{\lambda}(v)$. A possible null hypothesis is $H_0 : \Pr\{\lambda(v) = \mathit{const.}\} = 1$. Note that in this case, there is an additional difficulty of using existing results on specification test, as we have generated regressors $\hat{v}$.

Constructing a test of whether instruments are weak is an interesting topic but a more demanding one than above-mentioned tests. An inferential procedure that is robust to arbitrarily weak instruments also can be investigated.

# 11 Appendix

## 11.1 Proofs in Identification Analysis (Section 3.2)

In order to prove the sufficiency of Assumption ID2$'$ for ID2, we first introduce a preliminary lemma. For nonempty sets $A$ and $B$, define the following set

$$A+B=\{a+b:(a,b)\in A\times B\}. \tag{11.1}$$

Then, the following rules that are useful in proving the lemma. For nonempty sets $A$, $B$ and $C$,

$$\begin{aligned} A+B &= B+A \text{ (commutative)} && \text{(Rule 1)} \\ A+(B\cup C) &= (A+B)\cup(A+C) \text{ (distributive 1)} && \text{(Rule 2)} \\ A+(B\cap C) &= (A+B)\cap(A+C) \text{ (distributive 2)} && \text{(Rule 3)} \\ (A+B)^c &\subset A+B^c, && \text{(Rule 4)} \end{aligned}$$

where the last rule is less obvious than the others can be shown to hold. The distributive rules of Rule 2 and 3 do not hold when the operators are switched. For example, $A\cup(B+C)\neq(A\cup B)+(A\cup C)$.

Recall that $\mathcal{Z}_r=\left\{z\in\mathcal{Z}:rank\left(\frac{\partial\Pi(z)}{\partial z_2'}\right)=d_x\right\}$, and $\mathcal{Z}_0=\mathcal{Z}\backslash\mathcal{Z}_r$. We suppress the subscript "0" for the true functions for notational simplicity. Let $\lambda_{Leb}$ denotes a Lebesque measure on $\mathbb{R}^{d_x}$, and $\partial\mathcal{V}$ and $int(\mathcal{V})$ denote the boundary and interior of $\mathcal{V}$, respectively.

**Lemma 11.1** *Suppose Assumptions ID1 and ID2$'$(a)(i) and (ii) hold. Suppose $\mathcal{Z}_r\neq\phi$ and $\mathcal{Z}_0\neq\phi$. Then, (a) $\{\Pi(z)+v:z\in\mathcal{Z}_0,$ and $v\in int(\mathcal{V})\}\subset\mathcal{X}_r$, and (b) $\lambda_{Leb}(\Pi(\mathcal{Z}_0))=0$ and $\partial\mathcal{V}$ is countable.*

We prove this lemma after stating and proving the main lemma. The following lemma proves the sufficiency of Assumption ID2$'$:

**Lemma 11.2** *Suppose Assumption ID1 holds. Then, Assumption ID2$'$ implies Assumption ID2.*

In the following proofs, we distinguish the r.v.'s with their realization. Let $\xi$, $\zeta$, and $\nu$ denote the realizations of $x$, $z$, and $v$, respectively. Also, for expositional simplicity, we assume $z=z_2$ for the proofs of Lemmas 11.2 and 11.3.

**Proof of Lemma 11.2:** When $\mathcal{Z}_r = \phi$ or $\mathcal{Z}_r = \mathcal{Z}$ we trivially have $\mathcal{X}_r = \mathcal{X}$. Suppose $\mathcal{Z}_r \neq \phi$ and $\mathcal{Z}_0 \neq \phi$. First, under Assumption ID2$'$(b) that $\mathcal{V} = \mathbb{R}^{d_x}$, we have the conclusion by

$$\mathcal{X}_r = \left\{\Pi(z) + v : z \in \mathcal{Z}_r \text{ and } v \in \mathbb{R}^{d_x}\right\} = \mathbb{R}^{d_x} = \left\{\Pi(z) + v : x \in \mathcal{Z} \text{ and } v \in \mathbb{R}^{d_x}\right\} = \mathcal{X}.$$

Now suppose Assumption ID2$'$(a) holds. By Assumption ID2$'$(a)(iii), for $z \in \mathcal{Z}_0$, the joint support of $(z, v)$ is $\mathcal{Z}_0 \times \mathcal{V}$. Hence

$$\{\Pi(z) + v : z \in \mathcal{Z}_0, \text{ and } v \in int(\mathcal{V})\} = \{\Pi(z) + v : (z, v) \in \mathcal{Z}_0 \times int(\mathcal{V})\} = \Pi(\mathcal{Z}_0) + int(\mathcal{V}).$$

But by Lemma 11.1(a), $\Pi(\mathcal{Z}_0) + int(\mathcal{V}) \subset \mathcal{X}_r$ or contrapositively, $\mathcal{X}_r^c \subset (\Pi(\mathcal{Z}_0) + int(\mathcal{V}))^c$, and by (Rule 4), $(\Pi(\mathcal{Z}_0) + int(\mathcal{V}))^c \subset \Pi(Z_0) + \partial\mathcal{V}$. Therefore,

$$\mathcal{X} \backslash \mathcal{X}_r = \mathcal{X}_r^c \subset \Pi(Z_0) + \partial\mathcal{V}. \tag{11.2}$$

Let $\partial\mathcal{V} = \{\nu_1, \nu_2, ..., \nu_k, ...\} = \cup_{k=1}^{\infty}\{\nu_k\}$ by Lemma 11.1(b). Then

$$\begin{aligned}\lambda_{Leb}(\Pi(\mathcal{Z}_0) + \partial\mathcal{V}) &= \lambda_{Leb}(\Pi(\mathcal{Z}_0) + (\cup_{k=1}^{\infty}\{\nu_k\})) = \lambda_{Leb}(\cup_{k=1}^{\infty}(\Pi(\mathcal{Z}_0) + \{\nu_k\})) \\ &\leq \sum_{k=1}^{\infty}\lambda_{Leb}(\Pi(\mathcal{Z}_0) + \{\nu_k\}) = \sum_{k=1}^{\infty}\lambda_{Leb}(\Pi(\mathcal{Z}_0)) = 0,\end{aligned}$$

where the second equality is from (Rule 2) and the third equality by the property of Lebesgue measure. The last equality is by Lemma 11.1(b) that $\lambda_{Leb}(\Pi(\mathcal{Z}_0)) = 0$. Note that $x$ is a continuous r.v., and hence, by (11.2), $\Pr[x \in \mathcal{X}\backslash\mathcal{X}_r] \leq \Pr[x \in (\Pi(\mathcal{Z}_0)) + \partial\mathcal{V}] = 0$. □

**Proof of Lemma 11.1(a)**: First, we claim that for any $\pi \in \Pi(\mathcal{Z}_0)$ there exists $\cup_{n=1}^{\infty}\{\pi_n\} \subset \Pi(\mathcal{Z}_1)$ such that $\lim_{n\to\infty} \pi_n = \pi$.

By Proposition 4.21(a) of Lee (2011, p. 92), for any space $\mathcal{S}$, the path components of $\mathcal{S}$ form a partition of $\mathcal{S}$. Note that a path component of $\mathcal{S}$ is a maximal nonempty path connected subset of $\mathcal{S}$. Then, for $\mathcal{Z}_0 \subset \mathbb{R}^{d_z}$, we have $\mathcal{Z}_0 = \cup_{\iota\in I}\mathcal{Z}_{0\iota}$, where partitions $\mathcal{Z}_{0\iota}$ are path components. Note that, since $\mathcal{Z}_{0\iota}$ is path connected, for any $\zeta$ and $\tilde{\zeta}$ in $\mathcal{Z}_{0\iota}$, there exists a path in $\mathcal{Z}_{0\iota}$, namely, a piecewise continuously differentiable function $\gamma : [0, 1] \to \mathcal{Z}_{0\iota}$ such that $\gamma(0) = \zeta$ and $\gamma(1) = \tilde{\zeta}$. Note that $\{\gamma(t) : t \in [0, 1]\} \subset \mathcal{Z}_{0\iota}$. Consider a composite function $\Pi \circ \gamma : [0, 1] \to \Pi(\mathcal{Z}_{0\iota}) \subset \mathbb{R}^{d_x}$, then $\Pi(\gamma(\cdot))$ is differentiable, and by the mean value theorem, there exists $t^* \in [0, 1]$ such that

$$\Pi(\gamma(1)) - \Pi(\gamma(0)) = \frac{\partial\Pi(\gamma(t^*))}{\partial t}(1 - 0) = \frac{\partial\Pi(\gamma(t^*))}{\partial\zeta'}\frac{\partial\gamma(t^*)}{\partial t}.$$

Note that $\frac{\partial\Pi(\gamma(t^*))}{\partial\zeta_2'} = \mathbf{0}_{d_x\times d_x}$ since $\gamma(t^*) \in \mathcal{Z}_{0\iota} \subset \mathcal{Z}_0$ and $d_x = 1$. This implies that $\Pi(\gamma(1)) = \Pi(\gamma(0))$, or $\Pi(\zeta) = \Pi(\tilde{\zeta})$. Therefore for any $\zeta \in \mathcal{Z}_{0\iota}$,

$$\Pi(\zeta) = c_\iota \tag{11.3}$$

for some constant $c_\iota$ that depends on $\iota$.

Also, since $\mathcal{Z}_0$ is closed (see below), $\mathcal{Z}_{0\iota}$ is closed for any $\iota$. That is, $\mathcal{Z}_0$ is partitioned to a closed disjoint union of $\mathcal{Z}_{0\iota}$'s. But Assumption ID2$'$(a)(ii) says $\mathcal{Z}$ is a connected set in Euclidean space (i.e., $\mathbb{R}^{d_z}$). Therefore, for any $\iota \in I$, $\mathcal{Z}_{0\iota}$ must contain accumulation points of $\mathcal{Z}_r$ (Taylor (1985, p. 76)).

Now, for any $\pi = \Pi(\zeta) \in \Pi(\mathcal{Z}_0)$, it satisfies that $\zeta \in \mathcal{Z}_{0\iota}$ for some $\iota \in I$. Let $\zeta_c \in \mathcal{Z}_{0\iota}$ be an accumulation point of $\mathcal{Z}_r$, that is, there exists $\cup_{n=1}^{\infty}\{\zeta_n\} \subset \mathcal{Z}_r$ such that $\lim_{n\to\infty}\zeta_n = \zeta_c$. Then it follows that

$$\pi = \Pi(\zeta) = c_\iota = \Pi(\zeta_c) = \Pi(\lim_{n\to\infty}\zeta_n) = \lim_{n\to\infty}\Pi(\zeta_n) \equiv \pi_n,$$

where the second and third equalities are from (11.3) and the fourth by continuity of $\Pi(\cdot)$. Note that $\pi_n \in \Pi(\mathcal{Z}_r)$ for all $n \geq 1$. Therefore, we can conclude that for $\pi \in \Pi(\mathcal{Z}_0)$, there exists $\cup_{n=1}^{\infty}\{\pi_n\} \subset \Pi(\mathcal{Z}_r)$ such that $\lim_{n\to\infty}\pi_n = \pi$.

Now we prove that if $\xi \in \{\Pi(z) + v : z \in \mathcal{Z}_0, \text{ and } v \in int(\mathcal{V})\}$ then $\xi \in \mathcal{X}_r$. Suppose $\xi \in \{\Pi(z) + v : z \in \mathcal{Z}_0, \text{ and } v \in int(\mathcal{V})\}$, i.e., $\xi = \pi + \nu$ for $\pi \in \Pi(\mathcal{Z}_0)$ and $\nu \in int(\mathcal{V})$. Then, by the result above, there exists $\cup_{n=1}^{\infty}\{\pi_n\} \subset \Pi(\mathcal{Z}_r)$ such that $\lim_{n\to\infty}\pi_n = \pi$. Consider a sequence $\nu_n \equiv \xi - \pi_n$ for all $n \geq 1$. Notice that $\nu_n$ is not necessarily in $\mathcal{V}$. But,

$$\nu_n = (\pi + \nu) - \pi_n = \nu + (\pi - \pi_n),$$

hence $\lim_{n\to\infty}\nu_n = \nu$. Since $\nu \in int(\mathcal{V})$ there exists an open neighborhood of $\nu$, i.e., $B_\varepsilon(\nu)$ for some $\varepsilon$, such that $B_\varepsilon(\nu) \subset int(\mathcal{V})$. And, by the fact that $\lim_n \nu_n = \nu$, there exists $N_\varepsilon$ such that for all $n \geq N_\varepsilon$, $\nu_n \in B_\varepsilon(\nu)$. Therefore, by conveniently taking $n = N_\varepsilon$, it follows that

$$\xi = \pi_{N_\varepsilon} + \nu_{N_\varepsilon},$$

where $\pi_{N_\varepsilon} \in \Pi(\mathcal{Z}_r)$ and $\nu_{N_\varepsilon} \in B_\varepsilon(\nu) \in int(\mathcal{V}) \subset \mathcal{V}$. That is $\xi \in \mathcal{X}_r$. $\square$

**Proof that $\mathcal{Z}_0$ is closed in $\mathcal{Z}$**: Consider any $\cup_{n=1}^{\infty}\{\zeta_n\} \subset \mathcal{Z}_0 \subset \mathcal{Z}$ such that $\lim \zeta_n = \bar{\zeta}$. Then $\partial\Pi(\zeta_n)/\partial\zeta' = (0, 0, ..., 0) = \mathbf{0}$ by the definition of $\mathcal{Z}_0$, and

$$\partial\Pi(\bar{\zeta})/\partial\zeta' = \partial\Pi(\lim_{n\to\infty}\zeta_n)/\partial\zeta' = \lim_{n\to\infty}\partial\Pi(\zeta_n)/\partial\zeta' = \mathbf{0},$$

where the second equality is by continuity of $\partial\Pi(\cdot)/\partial\zeta'$. Therefore $rank(\partial\Pi_k(\bar{\zeta})/\partial\zeta') < d_x$ and $\bar{\zeta} \in \mathcal{Z}_0$, and hence $\mathcal{Z}_0$ is closed. $\square$

**Proof of Lemma 11.1(b)**: Recall $d_x = 1$. Note that $\mathcal{V} \subset \mathbb{R}$ can be expressed by a union of disjoint intervals. Since we are able to choose a rational number in each interval, the union is a countable union. But note that each interval has at most two end points which are the boundary of it. Therefore $\partial\mathcal{V}$ is countable. To prove that $\lambda_{Leb}(\Pi(\mathcal{Z}_0)) = 0$, we use the following proposition:

**Proposition 11.3 (Lusin-Saks, Corollary 6.1.3 in Garg (1998, p.92))** *Let $X$ be a* *normed vector space. Let $f : X \to \mathbb{R}$ and $E \subset X$. If at each point of $E$ at least one unilateral derivative of $f$ is zero, then $\lambda_{Leb}(f(E)) = 0$.*

Note that $\mathcal{Z}_0$ is the support where $\partial\Pi(z)/\partial z_k = 0$ for any $k \leq d_z$. Therefore, its bilateral (directional) derivative $D_\alpha \Pi(z)$ in the direction $\alpha = (\alpha_1, \alpha_2, ..., \alpha_{d_z})'$ satisfies $D_\alpha \Pi(z) = \sum_{k=1}^{d_z} \alpha_k \cdot \partial\Pi(z)/\partial z_k = 0$. Since the bilateral derivative is zero, each unilateral derivative is also zero; see, e.g., Giorgi, et al. (2004, p. 94) for the definitions of various derivatives. Then by Proposition 11.3, $\lambda_{Leb}(\Pi(\mathcal{Z}_0)) = 0$. $\square$

**Proof of Necessity Part of Theorem 3.5**: Suppose $\Pr(z \in \mathcal{Z}_1) = 0$. This implies $\Pr(z \in \mathcal{Z}_0) = 1$, but since $\mathcal{Z}_0$ is closed $\mathcal{Z}_0 = \mathcal{Z}$. Therefore, for any $z \in \mathcal{Z} = \mathcal{Z}_0$, the system of equations (3) either has multiple solutions or no solution. Therefore $g(\Pi(z) + v, z_1)$ is not identified. $\square$

## 11.2 Technical Assumptions and Proofs for Sufficiency (Section 6.1)

Assumptions B, C, D and L of Section 6.1 serve as sufficient conditions for more relevant technical assumptions that are directly used in the proofs for the convergence rate. In this section, we state Assumptions B.1, C.1 and D.1, and prove that these technical assumptions are implied by Assumptions B, C, D and L of the main body.

Let $\kappa = \kappa(n) = \min\{K_1 - 1, K_2 - 1\}$. Then $\kappa \asymp K$. Recall $Q = E[p^K(w_i)p^K(w_i)']$, where $p^K(w_i) = [p_-^{K_1-1}(x_i)' \vdots p^{K_2}(v_i)']'$ with $w_i = (x_i, v_i)$. As discussed in the main body, the transformation of the vector of regressors $p^K(w_i)$ produces the vector of new regressors $p^{*2\kappa+1}(u_i) = [1 \vdots \tilde{\Pi}(z_i)\partial p_-^\kappa(v_i)' \vdots p_-^\kappa(v_i)']'$ where $\tilde{\Pi}(\cdot) \in \mathcal{C}_1(\mathcal{Z})$ and $u_i = (z_i, v_i)$, and $Q^* = E[p^{*2\kappa+1}(u_i)p^{*2\kappa+1}(u_i)']$. Furthermore, since $\tilde{\Pi}(\cdot) \in \mathcal{C}_1(\mathcal{Z})$ can have nonempty $\mathcal{Z}_0$ as a subset of its domain, we define

$$\begin{aligned} Q_{z1}^* &= E[p^{*2\kappa+1}(u_i)p^{*2\kappa+1}(u_i)'|z_i \in \mathcal{Z}_r], \\ Q_{z0}^* &= E[p^{*2\kappa+1}(u_i)p^{*2\kappa+1}(u_i)'|z_i \in \mathcal{Z}_0]. \end{aligned}$$

Also define the second moment matrix for the first stage estimation as $Q_1 = E[r^L(z_i)r^L(z_i)']$. These matrices all depend on $n$. Also, for any matrix $A$, let the matrix norm be the Euclidean norm $\|A\| = \sqrt{tr(A'A)}$. And, for a symmetric matrix $B$, let $\lambda_{\min}(B)$ and $\lambda_{\max}(B)$ denote the minimum and maximum eigenvalues of $B$, respectively.

**Assumption B.1 (Bounds on second moment matrices)** (i) *$\lambda_{\min}(Q_{z1}^*)$ is bounded away from zero for all $\kappa(n)$, and $\lambda_{\min}(Q_1)$ is bounded away from zero for all $L(n)$, and*
(ii) *$\lambda_{\max}(Q^*)$ and $\lambda_{\max}(Q)$ are bounded by a fixed constant, for all $\kappa(n)$, and $\lambda_{\max}(Q_1)$ bounded by a fixed constant, for all $L(n)$.*

**Proof that B and L imply B.1:** Suppose $\Pr[z \in \mathcal{Z}_r(\tilde{\Pi})] = 1$, and it suffices to prove after replacing $Q_{z1}^*$ with $Q^*$ in Assumption B.1(i). Then $\tilde{\Pi}(\cdot)$ is piecewise one-to-one. Here, we prove the case where $\tilde{\Pi}(\cdot)$ is one-to-one, and the general case can be followed by conditioning on $z$ in each subset of $\mathcal{Z}$ where $\tilde{\Pi}(\cdot)$ is one-to-one.

Consider the change of variables of $u = (z, v)$ into $\tilde{u} = (\tilde{z}, \tilde{v})$ where $\tilde{z} = \tilde{\Pi}(z)$ and $\tilde{v} = v$. Then, it follows that $p^{*2\kappa+1}(u_i) = [1 \vdots \tilde{z}_i p_{-}^{\kappa}(v_i)' \vdots p_{-}^{\kappa}(v_i)']' = p^{2\kappa+1}(\tilde{u}_i)$ where $p^{2\kappa+1}(\tilde{u}_i)$ is one particular form of a vector of approximating functions as specified as in NPV (pp.572-573). Moreover, the joint density of $\tilde{u}$ is

$$f_{\tilde{u}}(\tilde{z}, \tilde{v}) = f_u(\tilde{\Pi}^{-1}(\tilde{z}), \tilde{v}) \cdot \begin{vmatrix} \frac{\partial \tilde{\Pi}^{-1}(\tilde{z})}{\partial \tilde{z}} & 0 \\ 0 & 1 \end{vmatrix} = f_u(\tilde{\Pi}^{-1}(\tilde{z}), \tilde{v}) \cdot \left| \frac{\partial \tilde{\Pi}^{-1}(\tilde{z})}{\partial \tilde{z}} \right|.$$

Since $\left| \frac{\partial \tilde{\Pi}^{-1}(\tilde{z})}{\partial \tilde{z}} \right| \neq 0$ by $\tilde{\Pi} \in \mathcal{C}_1(\mathcal{Z})$ (bounded derivative) and $f_u$ is bounded away from zero and the support of $u$ is compact by Assumption B, the support of $\tilde{u}$ is also compact and $f_{\tilde{u}}$ is also bounded away from zero. Then, by the proof of Theorem 4 in Newey (1997, p. 167), $\lambda_{\min}(E p^{2\kappa+1}(\tilde{u}_i) p^{2\kappa+1}(\tilde{u}_i)')$ is bounded away from zero. Therefore $\lambda_{\min}(Q^*)$ is bounded away from zero for all $\kappa$.

As for $Q_1$ that does not depend on the effect of weak instruments, the density of $z$ being bounded away from zero implies that $\lambda_{\min}(Q_1)$ is bounded away from zero for all $L$ by Newey (1997, p. 167) as above. The maximum eigenvalues of $Q^*$, $Q$ and $Q_1$ are bounded by fixed constants by the fact that the polynomials or splines are defined on bounded sets. □

**Assumption C.1 (Series approximation error)** *There exists* $\alpha, \alpha_1 > 0$, $\beta_{\tilde{K}}$ *and* $\gamma_{\tilde{L}}$ *such that* $\sup_{w \in \mathcal{W}} \left| h_0(w) - p^{\tilde{K}}(w)' \beta_{\tilde{K}} \right| \leq C \tilde{K}^{-\alpha}$ *as* $\tilde{K} \to \infty$, *and* $\sup_{z \in \mathcal{Z}} \left\| \Pi_0(z) - p^{\tilde{L}}(z)' \gamma_{\tilde{L}} \right\| \leq C \tilde{L}^{-\alpha_1}$ *as* $\tilde{L} \to \infty$.

This regulates the uniform approximation error of the series estimators $\hat{h}(\cdot)$ and $\hat{\Pi}(\cdot)$. As the orders of approximating functions grow, the errors shrink at the rates $\tilde{K}^{-\alpha}$ and $\tilde{L}^{-\alpha_1}$. Note that this is not affected by the weak instrument assumption.

**Proof that C implies C.1:** Take $\alpha = s/d_x$. Then, by Theorem 8 of Lorentz (1986, p. 90) for power series and by Theorem 12.8 of Schumaker (1981) for splines, differentiability of order $s$ guarantees Assumption C.1. Same applies to $\alpha_1 = s_1/d_z$. □

For the next assumption let $\zeta_r^v(\kappa)$ and $\xi_r^v(L)$ satisfy

$$\sup_{v \in \mathcal{V}} \left\| \partial^r p^{\kappa}(v) \right\| \leq \zeta_r^v(\kappa), \qquad \max_{|\mu| \leq r} \sup_{z \in \mathcal{Z}} \left\| \partial^{\mu} r^L(z) \right\| \leq \xi_r(L),$$

which impose nonstochastic uniform bounds on the vectors of approximating functions. Let $\Delta_\pi = \sqrt{L/n} + L^{-\alpha_1}$.

**Assumption D.1 (Rate of growth)** *For* $\kappa = \kappa(n)$ and $L = L(n)$, *the following convergence holds as* $n \to \infty$*:* (i) $n^{2\delta} \kappa^{1/2} \zeta_1^v(\kappa) \Delta_\pi \to 0$, *and* (ii) $n^{-\delta} \zeta_2^v(\kappa) \kappa^{1/2} \to 0$. *Also,* (iii) $\xi_0(L)^2 L/n \to 0$.

**Proof that D implies D.1:** It follows from Newey (1997, p. 157 and p. 160) that in the polynomial case

$$\zeta_r^v(K) = K^{1+2r} \tag{11.4}$$

and in the spline case that

$$\zeta_r^v(K) = K^{1/2+r}. \tag{11.5}$$

These are well-known properties of orthogonal polynomials and B-splines. The same results holds for $\xi_r(L)$. Therefore, Assumption D implies D.1. □

### 11.3 Key Lemmas for Proof of Rate of Convergence (Section 5.4)

To obtain the rate of convergence, a preliminary step is required to separate out the weak instrument factor as discussed in the main body. Lemmas 11.6 of this subsection describes that step and obtain the order of magnitudes of eigenvalues of the second moment matrices in term of the weak instrument rate. For ease of exposition, the proof the lemma will be given based on the case of $K_1 = K_2$ and univariate $x$. The general case of $K_1 \neq K_2$ and multivariate $x$ is discussed in Section 11.4 below. Recall $K = K_1 + K_2 - 1$ and we have $\kappa = \kappa(n) = K_1 = K_2 = (K+1)/2$. Again, $K_1$,$K_2$, $L$, $K$, and $\kappa$ all depends on $n$.

From the main body, we have

$$p^K(w_i)'T_n = p^{*K}(u_i)' + m_i^{K\prime},$$

where $p^{*K}(u_i)' = [1 \vdots \tilde{\Pi}(z_i)\partial p_-^{\kappa}(v_i)' \vdots p_-^{\kappa}(v_i)']$ and $m_i^{K\prime} = [0 \vdots \tilde{\Pi}(z_i)\left(\partial p_-^{\kappa}(\tilde{v}_i)' - \partial p_-^{\kappa}(v_i)'\right) \vdots (0_{\kappa\times 1})']$. Define

$$\tilde{Q} = \frac{P'P}{n},$$

where

$$\underset{n\times K}{P} = (p^K(w_1), ..., p^K(w_n))' \tag{11.6}$$

which is similar to (5.3) but with unobservable $v$ instead of residual $\hat{v}$. Note the distinction between $\tilde{Q}$ and $\hat{Q}$; recall $\hat{Q} = \frac{\hat{P}'\hat{P}}{n}$. Then, similar to (5.16)

$$T_n'\tilde{Q}T_n = \frac{T_n'P'PT_n}{n} = \frac{P^{*\prime}P^*}{n} + \frac{M'P^*}{n} + \frac{P^{*\prime}M}{n} + \frac{M'M}{n} \tag{11.7}$$

where $P^*$ and $M$ are matrices of $p^{*K}(u_i)$ and $m_i^K$, respectively, correspondingly defined as (11.6). Lastly, define the first term of (11.7) as

$$\tilde{Q}^* = \frac{P^{*\prime}P^*}{n}.$$

Also, recall $Q = E[p^K(w_i)p^K(w_i)']$ and $Q^* = E[p^{*K}(u_i)p^{*K}(u_i)']$.

In terms of the notations, it is useful to note that "$*$" for matrix $P$, $Q$, or vector $p^K(\cdot)$ which is free from the weak instrument effect due to the transformation. Also, "$\sim$" is for $P$ or

$Q$ with $v_i$'s in the arguments of approximating functions, and "$\hat{}$" is for the ones with $\hat{v}_i$'s in it. Before proceeding, we introduce two mathematical lemmas (Lemmas 11.4 and 11.5) that are useful in proving the main lemmas and theorem.

**Lemma 11.4** *For symmetric $k \times k$ matrices $A$ and $B$, let $\lambda_j(A)$ and $\lambda_j(B)$ denote their $j$th eigenvalues such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$. Then the following inequality holds: for $1 \leq j \leq k$,*

$$|\lambda_j(A) - \lambda_j(B)| \leq |\lambda_1(A - B)| \leq \|A - B\| .$$

By having $i = 1$ and $n$, note that Lemma 11.4 implies $|\lambda_{\max}(A) - \lambda_{\max}(B)| \leq \|A - B\|$ and $|\lambda_{\min}(A) - \lambda_{\min}(B)| \leq \|A - B\|$, respectively, which will be useful in several proofs below.

**Proof of Lemma 11.4:** We provide slightly more general results. Firstly, by Weyl (1912), for symmetric $k \times k$ matrices $C$ and $D$

$$\lambda_{i+j-1}(C + D) \leq \lambda_i(C) + \lambda_j(D), \tag{11.8}$$

where $i + j - 1 \leq k$. As for the second inequality, we prove

$$|\lambda_j(D)| \leq \|D\| , \tag{11.9}$$

for $1 \leq j \leq k$. Note that, for any $k \times 1$ vector $a$ such that $\|a\| = 1$,

$$(a'Da)^2 = tr(a'Daa'Da) = tr(DDaa'aa') = tr(DDaa') \leq tr(DD)tr(aa') = tr(DD).$$

Since $\lambda_j(D) = a'Da$ for some $a$ with $\|a\| = 1$, taking square root on both sides of the inequality gives the desired result. Now, in inequalities (11.8) and (11.9), take $j = 1$, $C = B$, and $D = A - B$ and we have the conclusions. □

**Lemma 11.5** *If $K(n) \times K(n)$ symmetric random sequence of matrices $A_n$ satisfies $\lambda_{\max}(A_n) = O_p(n^\delta)$, then $\|B_n A_n\| \leq \|B_n\| O_p(n^\delta)$ for a given sequence of matrices $B_n$.*

Another useful corollary of this lemma is that, for $\lambda_{\max}(A_n) = O_p(n^\delta)$ and sequence of vectors $b_n$ and $c_n$, we have $b_n' A_n c_n \leq O_p(n^\delta) b_n' c_n$.

**Proof of Lemma 11.5:** Let $A_n$ have eigenvalue decomposition $A_n = UDU^{-1}$. Then

$$\begin{aligned}
\|B_n A_n\|^2 &= tr\left(B_n A_n A_n B_n'\right) = tr\left(B_n UDU^{-1}UDU^{-1}B_n'\right) \\
&= tr\left(B_n UD^2U^{-1}B_n'\right) \leq tr\left(B_n UU^{-1}B_n'\right) \cdot \lambda_{\max}(A_n)^2 \\
&= \|B_n\|^2 O_p(n^\delta)^2.
\end{aligned}$$

□

The following lemma is the main result of this subsection.

**Lemma 11.6** *Suppose Assumptions ID, A, B.1, D.1, and L are satisfied. Then, (a)*

$$\lambda_{\max}(Q^{-1}) = O(n^{2\delta}),$$

*(b)*

$$\lambda_{\max}(\hat{Q}^{-1}) = O_p(n^{2\delta}).$$

In all proofs, let $C$ denote a generic positive constant that may be different in different use. TR, CS, MK are triangular inequality, Cauchy-Schwartz inequality and Markov inequality, respectively. Also, "$w.p. \to 1$" stands for "with probability approaching one."

**Preliminary derivations for the proofs of Lemmas 11.6-7 and Theorem 6.1:** Before proving the lemmas and theorems below, it is useful to list the implications of Assumption D.1(i) that are used in the proofs. Define

$$\Delta_\pi = \sqrt{L/n} + L^{-\alpha_1}, \quad \Delta_{\hat{Q}} = \zeta_1^v(\kappa)^2 \Delta_\pi^2 + \kappa^{1/2}\zeta_1^v(\kappa)\Delta_\pi, \quad \Delta_{\tilde{Q}} = \sqrt{\zeta_1^v(\kappa)^2 \kappa/n}.$$

Note that $\Delta_{\tilde{Q}} = \zeta_1^v(\kappa)\sqrt{\kappa/n} \to 0$ by

$$n^{2\delta}\kappa^{1/2}\zeta_1^v(\kappa)\Delta_\pi \to 0 \tag{11.10}$$

of Assumption D.1(i). Also

$$n^{2\delta}\Delta_{\hat{Q}} = n^{2\delta}\left\{\zeta_1^v(\kappa)^2\Delta_\pi^2 + \kappa^{1/2}\zeta_1^v(\kappa)\Delta_\pi\right\} \to 0$$
$$n^{2\delta}\Delta_{\tilde{Q}} = n^{2\delta}\sqrt{\zeta_1^v(\kappa)^2\kappa/n} \le Cn^{2\delta}\kappa^{1/2}\zeta_1^v(\kappa)/\sqrt{n} \to 0$$

and

$$n^{2\delta}\zeta_1^v(\kappa)^2\Delta_\pi^2/n \to 0 \tag{11.11}$$

by $n^{\delta}\kappa^{1/2}\zeta_1^v(\kappa)\Delta_\pi \to 0$ and $n^{2\delta}\zeta_1^v(\kappa)\Delta_\pi \to 0$ which are implied by (11.10).

**Proof of Lemma 11.6(a):** Let $p_i^* = p^{*K}(u_i)$ and $m_i = m_i^K$ for brevity. Recall (5.16) that $T_n' Q T_n = Q^* + E\left[m_i p_i^{*\prime}\right] + E\left[p_i^* m_i'\right] + E\left[m_i m_i'\right]$. Then

$$\left\|T_{sn}' Q T_{sn} - Q^*\right\| \le 2E\left\|m_i\right\|\left\|p_i^*\right\| + E\left\|m_i\right\|^2 \le 2\left(E\left\|m_i\right\|^2\right)^{1/2}\left(E\left\|p_i^*\right\|^2\right)^{1/2} + E\left\|m_i\right\|^2$$

by Cauchy-Schwartz inequality. But $m_i = m_i^K = [0 \vdots \tilde{\Pi}(z_i)\left(\partial p_-^\kappa(\tilde{v}_i)' - \partial p_-^\kappa(v_i)'\right) \vdots (0_{\kappa\times 1})']'$ where $\tilde{v}$ is the intermediate value between $x$ and $v$. Then, by mean value expanding $\partial p^\kappa(\tilde{v}_i)$ around $v_i$ and $|\tilde{v}_i - v_i| \le |x_i - v_i|$, we have

$$\begin{aligned} \left\|m_i\right\|^2 &= \left\|\tilde{\Pi}(z_i)\partial^2 p_-^\kappa(\bar{v}_i)\left(\tilde{v}_i - v_i\right)\right\|^2 \le \left|\tilde{\Pi}(z_i)\right|^2 \zeta_2^v(\kappa)^2 \left|x_i - v_i\right|^2 \\ &= n^{-2\delta}\left|\tilde{\Pi}(z_i)\right|^4 \zeta_2^v(\kappa)^2 \le Cn^{-2\delta}\zeta_2^v(\kappa)^2, \end{aligned} \tag{11.12}$$

where $\bar{v}$ is the intermediate value between $v$ and $\tilde{v}$, and by Assumption L that $\sup_z\left|\tilde{\Pi}(z_i)\right| <$

$\infty$. Therefore

$$E\left\|m_i\right\|^2 \leq Cn^{-2\delta}\zeta_2^v(\kappa)^2. \tag{11.13}$$

Then, by Assumption B.1(ii),

$$E\left[p_i^{*\prime}p_i^*\right] = tr(Q^*) \leq tr(I_{2\kappa})\lambda_{\max}(Q^*) \leq C\cdot\kappa. \tag{11.14}$$

Therefore, by combining (11.13) and (11.14) it follows

$$\left\|T_n'QT_n - Q^*\right\| \leq O(\kappa^{1/2}n^{-2\delta}\zeta_2^v(\kappa)) + O(n^{-2\delta}\zeta_2^v(\kappa)^2) = o(1) \tag{11.15}$$

by Assumption D.1(ii), which shows that all the remainder terms are negligible.

Now, by Lemma 11.4, we have

$$\left|\lambda_{\min}(T_n'QT_n) - \lambda_{\min}(Q^*)\right| \leq \left\|T_n'QT_n - Q^*\right\| \tag{11.16}$$

Combine the results (11.15) and (11.16) to have $\lambda_{\min}(T_n'QT_n) = \lambda_{\min}(Q^*) + o(1)$. But note that, with simpler notations $p_{z1} = \Pr[z \in \mathcal{Z}_r]$ and $p_{z0} = \Pr[z \in \mathcal{Z}_0]$, we can write $Q^* = p_{z1}Q_{z1}^* + p_{z0}Q_{z0}^*$. Then, by a variant of Lemma 11.4 (with the fact that $\lambda_1(-B) = -\lambda_k(B)$ for any symmetric matrix $B$), it follows that $\lambda_{\min}(Q^*) \geq p_{z1}\cdot\lambda_{\min}(Q_{z1}^*) + p_{z0}\cdot\lambda_{\min}(Q_{z0}^*) = p_{z1}\cdot\lambda_{\min}(Q_{z1}^*)$, as $\lambda_{\min}(Q_{z0}^*) = 0$. Then, since $p_{z1} > 0$, it holds that $\lambda_{\min}(Q^*) \geq p_{z1}\cdot\lambda_{\min}(Q_{z1}^*) \geq c > 0$ for all $\kappa(n)$ by Assumption B.1(i). Therefore,

$$\lambda_{\min}(T_n'QT_n) \geq c > 0. \tag{11.17}$$

Let

$$T_{0n} = \begin{bmatrix} n^\delta & 0 \\ -n^\delta & 1 \end{bmatrix} \otimes I_\kappa, \quad \text{so that} \quad T_n = \begin{bmatrix} 1 & 0_{1\times 2\kappa} \\ 0_{2\kappa\times 1} & T_{0n} \end{bmatrix}.$$

Then, by solving $\begin{vmatrix} n^\delta - \tilde{\lambda} & 0 \\ -n^\delta & 1-\tilde{\lambda} \end{vmatrix} = 0$, we have $\tilde{\lambda} = n^\delta$ or 1 for eigenvalues of $T_{0n}$, and since $\lambda_{\max}(I_\kappa) = 1$, it follows

$$\lambda_{\max}(T_n) = \lambda_{\max}(T_{0n}) = O(n^\delta). \tag{11.18}$$

Note that $\lambda_{\max}(T_nT_n') = O(n^{2\delta})$ by Lemma 11.5. Since (11.17) implies $\lambda_{\max}((T_n'QT_n)^{-1}) = O(1)$, it follows

$$\lambda_{\max}(Q^{-1}) = \lambda_{\max}(T_n(T_n'QT_n)^{-1}T_n') \leq O(1)\lambda_{\max}(T_nT_n') = O(n^{2\delta})$$

by applying Lemma 11.5 again. □

**Proof of Lemma 11.6(b):** For notational simplicity let $p_i = p^K(w_i)$ whose element is denoted as $p_k(w_i)$ for $k = 1, ..., K$. Consider

$$\begin{aligned} E\left\|\tilde{Q}-Q\right\|^2 &= E\left\{\sum_j\sum_k\left(\left[\frac{\sum_i p_i p_i'}{n} - E\left[p_i p_i'\right]\right]_{jk}\right)^2\right\} \\ &= \sum_j\sum_k E\left\{\left(\frac{\sum_i p_k(w_i)p_j(w_i)}{n} - E\left[p_k(w_i)p_j(w_i)\right]\right)^2\right\} \\ &\leq \sum_j\sum_k \frac{1}{n}E\left[p_k(w_i)^2 p_j(w_i)^2\right] = \frac{1}{n}E\left[p_i' p_i p_i' p_i\right] \end{aligned}$$

by Assumptions A (first ineq.). Then, we bound the forth moment with the bounds of a second moment and the following bound. With $w = (x, v)$,

$$\max_{|\mu|\leq r}\sup_{w\in\mathcal{W}}\left\|\partial^\mu p^K(w)\right\|^2 \leq \sup_{x\in\mathcal{X}}\left\|\partial^r p^\kappa(x)\right\|^2 + \sup_{v\in\mathcal{V}}\left\|\partial^r p^\kappa(v)\right\|^2 \leq \zeta_r^v(\kappa)^2 + \zeta_r^v(\kappa)^2 = 2\zeta_r^v(\kappa)^2.$$

By Assumption B.1(ii), the second moment $E\left[p_i' p_i\right] = tr(Q) \leq tr(I_{2\kappa})\lambda_{\max}(Q) \leq C\cdot\kappa$. Hence

$$E\left\|\tilde{Q}-Q\right\|^2 \leq \frac{1}{n}E\left[p_i' p_i p_i' p_i\right] \leq O(\zeta_1^v(\kappa)^2\kappa/n).$$

Therefore, by MK, $\left\|\tilde{Q}-Q\right\|^2 = O_p(\zeta_1^v(\kappa)^2\kappa/n) = O_p(\Delta_{\tilde{Q}}^2)$. Now

$$\left\|T_n'\tilde{Q}T_n - T_n'QT_n\right\| = \left\|T_n'(\tilde{Q}-Q)T_n\right\| \leq \lambda_{\max}(T_n)^2\left\|\tilde{Q}-Q\right\| \leq O(n^{2\delta_1})O_p(\Delta_{\tilde{Q}}) = o_p(1) \tag{11.19}$$

by Assumption D.1(i) and (11.18).

Let $\hat{p}_i = p^K(\hat{w}_i) = [1 \vdots p_-^\kappa(x_i) \vdots p_-^\kappa(\hat{v}_i)]$ and by mean value expansion

$$\hat{p}_i = [1 \vdots p_-^\kappa(x_i) \vdots p_-^\kappa(v_i) + \partial p_-^\kappa(\bar{v}_i)\,(\hat{v}_i - v_i)],$$

where $\bar{v}$ is the intermediate value. Since

$$\frac{1}{n}\sum_i\left|\hat{v}_i - v\right|^2 = \frac{1}{n}\sum_i\left|\Pi(z_i) - \hat{\Pi}(z_i)\right|^2 = O_p(\Delta_\pi^2), \tag{11.20}$$

we have

$$\begin{aligned} \left\|\hat{p}_i - p_i\right\|^2 &= \left\|p_-^\kappa(x_i) - p_-^\kappa(x_i)\right\|^2 + \left\|\tilde{\Pi}(z_i)\partial p_-^\kappa(\bar{v}_i)\,(\hat{v}_i - v_i)\right\|^2 \\ &\leq C\zeta_1^v(\kappa)^2\frac{1}{n}\sum_i\left|\hat{v}_i - v_i\right|^2 \leq O_p(\zeta_1^v(\kappa)^2\Delta_\pi^2). \end{aligned} \tag{11.21}$$

Also, by MK,

$$\Pr[\left\|p_i\right\|^2 > \varepsilon] \leq \frac{E\left\|p_i\right\|^2}{\varepsilon} = C\cdot tr(Q) \leq C\cdot tr(I_{2\kappa})\lambda_{\max}(Q) = O(\kappa), \tag{11.22}$$

hence $\sum_{i=1}^{n}\|p_i\|^2/n = O_p(\kappa)$. Then, by TR, CS and by combining (11.21) and (11.22)

$$\begin{aligned}
\left\|\hat{Q}-\tilde{Q}\right\| &= \left\|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{p}_i\hat{p}_i'-p_ip_i'\right)\right\| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left\|\left(\hat{p}_i-p_i\right)\left(\hat{p}_i-p_i\right)'+p_i(\hat{p}_i-p_i)'+(\hat{p}_i-p_i)p_i'\right\| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\|\hat{p}_i-p_i\|^2+2\frac{1}{n}\sum_{i=1}^{n}\|p_i\|\,\|\hat{p}_i-p_i\| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\|\hat{p}_i-p_i\|^2+2\left(\frac{1}{n}\sum_{i=1}^{n}\|p_i\|^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\|\hat{p}_i-p_i\|^2\right)^{1/2} \\
&= O_p(\zeta_1^v(\kappa)^2\Delta_\pi^2)+O_p(\kappa^{1/2})O_p(\zeta_1^v(\kappa)\Delta_\pi)=O_p(\Delta_{\hat{Q}}).
\end{aligned}$$

Thus, by Lemma 11.5,

$$\left\|T_n'\hat{Q}T_n-T_n'\tilde{Q}T_n\right\|=\left\|T_n'(\hat{Q}-\tilde{Q})T_n\right\|\leq O(n^{2\delta})O_p(\Delta_{\hat{Q}})=o_p(1) \tag{11.23}$$

by Assumption D.1(i) and (11.18). Therefore, by combining (11.19) and (11.23) and by TR,

$$\left\|T_n'\hat{Q}T_n-T_n'QT_n\right\|\leq\left\|T_n'\hat{Q}T_n-T_n'\tilde{Q}T_n\right\|+\left\|T_n'\tilde{Q}T_n-T_n'QT_n\right\|=o_p(1). \tag{11.24}$$

Now by TR, (11.24) and (11.15) give $\left\|T_n'\hat{Q}T_n-Q^*\right\|\leq\left\|T_n'\hat{Q}T_n-T_n'QT_n\right\|+\|T_n'QT_n-Q^*\|$ $=o_p(1)$. Also, by Lemma 11.4, we have $\left|\lambda_{\min}(T_n'\hat{Q}T_n)-\lambda_{\min}(Q^*)\right|\leq\left\|T_n'\hat{Q}T_n-Q^*\right\|$. Combine the results to have $\lambda_{\min}(T_n'\hat{Q}T_n)=\lambda_{\min}(Q^*)+o_p(1)$. But, by Assumption B.1(i), we have

$$\lambda_{\min}(T_n'\hat{Q}T_n)\geq c>0, \tag{11.25}$$

$w.p\to 1$ as $n\to\infty$. Then, similarly to the proof of Lemma 11.6(a), we have

$$\lambda_{\max}(\hat{Q}^{-1})=\lambda_{\max}(T_n(T_n'\hat{Q}T_n)^{-1}T_n')\leq O_p(1)\lambda_{\max}(T_nT_n')=O_p(n^{2\delta}).$$

□

## 11.4 Generalization

The general case of $K_1\neq K_2$ for Section 11.3 can be incorporated by the following argument. Recall that $K=K_1+K_2-1$ and $\kappa=\min\{K_1-1,K_2-1\}$. We assume $K_1\asymp K_2$ then

$K \asymp \kappa$. Suppose $K_1 \leq K_2$, then we can rewrite $p^K(w)$ as

$$\begin{aligned} p^K(w) &= (1, p_2(x), ..., p_{K_1}(x), p_2(v), ..., p_{K_2}(v))' \\ &= (1, p_2(x), ..., p_{K_1}(x), p_2(v), ..., p_{K_1}(v), p_{K_1+1}(v), ..., p_{K_2}(v))' \\ &= \left[1 \vdots p_-^{K_1}(x)' \vdots p_-^{K_1}(v)' \vdots (p_{K_1+1}(v), ..., p_{K_2}(v))'\right]' \\ &= \left[p^{2K_1-1}(w)' \vdots d^{K_2-K_1}(v)'\right]'. \end{aligned}$$

where a vector $d^{K_2-K_1}(v) = (p_{K_1+1}(v), ..., p_{K_2}(v))'$. Let $p^{2\kappa-1}(w_i) = \tilde{p}_i$ and $d^{K_2-K_1}(v_i) = d_i$. Then

$$Q = E[p^K(w)p^K(w)'] = \begin{bmatrix} E[\tilde{p}_i\tilde{p}_i'] & E[\tilde{p}_i d_i'] \\ E[d_i\tilde{p}_i'] & E[d_i d_i'] \end{bmatrix}$$

Note that $E[d_i d_i']$ is the usual second moment matrix free from the weak instruments. Therefore, the singularity of $Q$ is determined by the remaining three matrices. Note that $\tilde{p}_i$ is subject to the weak instruments as before and after mean value expanding it, the usual transformation matrix can be applied. Then by applying the inverse of partitioned matrix, the maximum eigenvalue of $Q^{-1}$ can eventually be expressed in term of the $n^\delta$ rate of Assumption L. A similar argument holds for $K_1 \geq K_2$.

We can also have a more general case of multivariate $x$ for the proof of Lemma 11.6 in Section 11.3 and the derivation in Section 5.4 of the main body. Define

$$\partial^\mu p_j(x) = \frac{\partial^{|\mu|} p_j(x)}{\partial x_1^{\mu_1} \partial x_2^{\mu_2} \cdots \partial x_{d_x}^{\mu_{d_x}}}$$

where $|\mu| = \sum_{t=1}^{d_x} \mu_t$, and $\partial^\mu p^K(x) = [\partial^\mu p_1(x), ..., \partial^\mu p_{K_2}(x)]'$. Then, multivariate mean value expansion can be applied using the derivative $\partial^\mu p_j(x)$. Whether instruments that are associated with $x$ have different strengths (hence different rates) or not does not change the strategy. It will become similar to a proof with a linear reduced form with different strength of instruments.

One can also have a simpler proof by considering mean value expansion of just one element of $x$ and the similar proof follows as the univariate case. That is, $p_j(v) = p_j(x - \Pi_n(z)) = p_j(x) - \Pi_{m,n}(z)\partial p_j(\tilde{x})/\partial x_m$ where $\tilde{x}$ is an intermediate value and $\Pi_{m,n}(\cdot)$ is $m$-th element of $\Pi_n(\cdot)$.

## 11.5 Proofs of Rate of Convergence (Section 6.2)

Given the results of the lemmas above, we prove the rate of convergence. We first prove a lemma with the unpenalized series estimator $\hat{h}(\cdot)$ defined in Section 5.1, and then prove the main theorem with the penalized estimator $\hat{h}_\tau(\cdot)$ defined in Section 5.3.

**Lemma 11.7** *Suppose Assumptions A-D, and L are satisfied. Then,*

$$\left\|\hat{h}-h_0\right\|_{L_2} = O_p\left(n^{\delta}(\sqrt{K/n}+K^{-s/d_x}+\sqrt{L/n}+L^{-s_1/d_z})\right).$$

**Proof of Lemma 11.7:** Let $\beta$ be such that $\sup_{\mathcal{W}}\left|h_0(w)-p^K(w)'\beta\right| = O(K^{-\alpha})$ (by Assumption C.1). Then, by TR of $L_2$ norm (first ineq.),

$$\begin{aligned}
\left\|\hat{h}-h_0\right\|_{L_2} &= \left\{\int\left[\hat{h}(w)-h_0(w)\right]^2 dF(w)\right\}^{1/2}\\
&\leq \left\{\int\left[p^K(w)'(\hat{\beta}-\beta)\right]^2 dF(w)\right\}^{1/2}+\left\{\int\left[p^K(w)'\beta-h_0(w)\right]^2 dF(w)\right\}^{1/2}\\
&= \left\{\left(\hat{\beta}-\beta\right)' Ep^K(w)p^K(w)'\left(\hat{\beta}-\beta\right)\right\}^{1/2}+O(K^{-\alpha})\\
&\leq C\left\|\hat{\beta}-\beta\right\|+O(K^{-\alpha})
\end{aligned}$$

by Assumption B.1(ii) and Lemma 11.5 (last eq.). As $\hat{\beta}-\beta=(\hat{P}'\hat{P})^{-1}\hat{P}'(y-\hat{P}\beta)$, it follows that

$$\begin{aligned}
\left\|\hat{\beta}-\beta\right\|^2 &= (\hat{\beta}-\beta)'(\hat{\beta}-\beta)\\
&= (y-\hat{P}\beta)'\hat{P}(\hat{P}'\hat{P})^{-1}(\hat{P}'\hat{P})^{-1}\hat{P}'(y-\hat{P}\beta)\\
&= (y-\hat{P}\beta)'\hat{P}\hat{Q}^{-1}\hat{Q}^{-1}\hat{P}'(y-\hat{P}\beta)/n^2\\
&= (y-\hat{P}\beta)'\hat{P}\hat{Q}^{-1/2}\hat{Q}^{-1}\hat{Q}^{-1/2}\hat{P}'(y-\hat{P}\beta)/n^2\\
&\leq O_p(n^{2\delta})(y-\hat{P}\beta)'\hat{P}\left(\hat{P}'\hat{P}\right)^{-1}\hat{P}'(y-\hat{P}\beta)/n
\end{aligned}$$

by Lemma 11.5 and Lemma 11.6(b) (last ineq.).

Let $h=(h(w_1),...,h(w_n))'$ and $\tilde{h}=(h(\hat{w}_1),...,h(\hat{w}_n))'$. Also let $\eta_i=y_i-h_0(w_i)$ and $\eta=(\eta_1,...,\eta_n)'$. Let $W=(w_1,..,w_n)'$, then $E[y_i|W]=h_0(w_i)$ which implies $E[\eta_i|W]=0$. Also similar to the proof of Lemma A1 in NPV (p.594), by Assumption A, we have $E[\eta_i^2|W]$ being bounded and $E[\eta_i\eta_j|W]=0$ for $i\neq j$. (Here, the expectation is taken for $y$.) Then, given that $y-\hat{P}\beta=(y-h)+(h-\tilde{h})+(\tilde{h}-\hat{P}\beta)$, we have, by TR,

$$\begin{aligned}
\left\|\hat{\beta}-\beta\right\| &= O_p(n^{\delta})\left\|\hat{Q}^{-1/2}\hat{P}'(y-\hat{P}\beta)/n\right\|\\
&\leq O_p(n^{\delta})\{\left\|\hat{Q}^{-1/2}\hat{P}'\eta/n\right\|+\left\|\hat{Q}^{-1/2}\hat{P}'(h-\tilde{h})/n\right\|\\
&\quad+\left\|\hat{Q}^{-1/2}\hat{P}'(\tilde{h}-\hat{P}\beta)/n\right\|\}. \qquad (11.26)
\end{aligned}$$

For the first term of equation (11.26), consider

$$\begin{aligned}
E\left[\left\|(PT_n-P^*)'\eta/n\right\|^2\middle|W\right] &= E\left[\left\|M'\eta/n\right\|^2\middle|W\right]\leq C\frac{1}{n^2}\sum_i\left\|m_i\right\|^2\\
&= O_p(n^{-2\delta-1}\zeta_2^v(\kappa)^2)=o_p(1)
\end{aligned}$$

by (11.12) and $o_p(1)$ is implied by Assumption D.1(ii). Therefore, by MK,

$$\left\|(PT_n - P^*)'\eta/n\right\| = o_p(1). \tag{11.27}$$

Also,

$$\begin{aligned} E\left[\left\|\left(\hat{P}T_n - PT_n\right)'\eta/n\right\|^2 \middle| W\right] &\leq C\frac{1}{n^2}\sum_i \left\|(\hat{p}_i - p_i)'T_n\right\|^2 \leq C\frac{1}{n^2}\sum_i \lambda_{\max}(T_n)^2 \left\|\hat{p}_i - p_i\right\|^2 \\ &\leq \frac{1}{n}O(n^{2\delta})O_p(\zeta_1^v(\kappa)^2\Delta_\pi^2) = O_p(n^{2\delta}\zeta_1^v(\kappa)^2\Delta_\pi^2/n) \end{aligned} \tag{11.28}$$

by (11.18) and (11.21), and hence

$$\left\|\left(\hat{P}T_n - PT_n\right)'\eta/n\right\| = o_p(1) \tag{11.29}$$

by Assumption D.1(i) (or (11.11)) and MK. Also

$$\begin{aligned} E\left\|P^{*\prime}\eta/n\right\|^2 &= E\left[E[\left\|P^{*\prime}\eta/n\right\|^2|W]\right] = E\left[\sum_i p_i^{*\prime}p_i^* E[\eta_i^2|W]/n^2\right] \\ &\leq C\frac{1}{n^2}\sum_i E\left[p_i^{*\prime}p_i^*\right] = Ctr(Q^*)/n = O(\kappa/n) \end{aligned}$$

by Assumptions A (first ineq.), and equation (11.14) (last eq.). By MK, this implies

$$\left\|P^{*\prime}\eta/n\right\| \leq O_p(\sqrt{\kappa/n}). \tag{11.30}$$

Hence by TR with (11.27), (11.29), and (11.30),

$$\left\|T_n'\hat{P}'\eta/n\right\| \leq \left\|\left(\hat{P}T_n - PT_n\right)'\eta/n\right\| + \left\|(PT_n - P^*)'\eta/n\right\| + \left\|P^{*\prime}\eta/n\right\| \leq O_p(\sqrt{\kappa/n}).$$

Therefore, the first term of (11.26) becomes

$$\left\|\hat{Q}^{-1/2}\hat{P}'\eta/n\right\|^2 = \left(\frac{\eta'\hat{P}T_n}{n}\right)(T_n'\hat{Q}T_n)^{-1}\left(\frac{T_n'\hat{P}'\eta}{n}\right) \leq O_p(1)\left\|T_n'\hat{P}'\eta/n\right\|^2 = O_p(\kappa/n) \tag{11.31}$$

by Lemma 11.5 and (11.25).

Due to the fact that $I - \hat{P}\left(\hat{P}'\hat{P}\right)^{-1}\hat{P}'$ is a projection matrix, hence is p.s.d, the second term of (11.26) becomes

$$\begin{aligned} \left\|\hat{Q}^{-1/2}\hat{P}'(h - \tilde{h})/n\right\|^2 &= (h - \tilde{h})'\hat{P}\left(\hat{P}'\hat{P}\right)^{-1}\hat{P}'(h - \tilde{h})/n \leq (h - \tilde{h})'(h - \tilde{h})/n \\ &= \sum_i \left(h(w_i) - h(\hat{w}_i)\right)^2/n = \sum_i \left(\lambda(v_i) - \lambda(\hat{v}_i)\right)^2/n \\ &\leq C\sum_i \left|v_i - \hat{v}_i\right|^2/n = O_p(\Delta_\pi^2) \end{aligned} \tag{11.32}$$

by (11.20) (last eq.) and Assumption C (Lipschitz continuity of $\lambda(v)$) (last ineq.). This term is due to the generated regressors $\hat{v}$ from the first stage estimation, and hence follows the rate of the first stage series estimation ($\Delta_\pi^2$). Similarly, the last term is

$$\begin{aligned}\left\|\hat{Q}^{-1/2}\hat{P}'(\tilde{h}-\hat{P}\beta)/n\right\|^2 &= (\tilde{h}-\hat{P}\beta)'\hat{P}\left(\hat{P}'\hat{P}\right)^{-1}\hat{P}'(\tilde{h}-\hat{P}\beta)/n \\ &\leq (\tilde{h}-\hat{P}\beta)'(\tilde{h}-\hat{P}\beta)/n \\ &= \sum_i \left(h(\hat{w}_i)-p^K(\hat{w}_i)'\beta\right)^2/n = O_p(K^{-2\alpha}) \qquad (11.33)\end{aligned}$$

by Assumption C.1. Therefore, by combining (11.31), (11.32), and (11.33)

$$\left\|\hat{\beta}-\beta\right\| \leq O_p(n^\delta)\left[O_p(\sqrt{\kappa/n})+O_p(\Delta_\pi)+O(K^{-\alpha})\right].$$

Consequently, since $\kappa \asymp K$,

$$\left\|\hat{h}-h_0\right\|_{L_2} \leq O_p(n^\delta)\left[O_p(\sqrt{K/n})+O(K^{-\alpha})+O_p(\Delta_\pi)\right]+O(K^{-\alpha})$$

and we have the conclusion of the lemma. □

**Proof of Theorem 6.1:** Now we derive convergence rate of the penalized series estimator $\hat{h}_\tau(\cdot)$. Recall $\hat{Q}_\tau = (\hat{P}'\hat{P}+n\tau_n I)/n = \hat{Q}+\tau_n I$. Define $\hat{P}_\# = \hat{P}+n\tau_n\hat{P}(\hat{P}'\hat{P})^{-1}$. Note that the penalty bias emerges as $\hat{P}_\# \neq \hat{P}$. Consider

$$\begin{aligned}\left\|\hat{\beta}_\tau-\beta\right\|^2 &= (\hat{\beta}_\tau-\beta)'(\hat{\beta}_\tau-\beta) \\ &= (y-\hat{P}_\#\beta)'\hat{P}(\hat{P}'\hat{P}+n\tau_n I)^{-1}(\hat{P}'\hat{P}+n\tau_n I)^{-1}\hat{P}'(y-\hat{P}_\#\beta) \\ &= (y-\hat{P}_\#\beta)'\hat{P}\hat{Q}_\tau^{-1/2}\hat{Q}_\tau^{-1}\hat{Q}_\tau^{-1/2}\hat{P}'(y-\hat{P}_\#\beta)/n^2 \\ &\leq \lambda_{\max}(\hat{Q}_\tau^{-1})\left\|\hat{Q}_\tau^{-1/2}\hat{P}'(y-\hat{P}_\#\beta)/n\right\|^2\end{aligned}$$

Then, first note that, by (5.11) and Lemma 11.6(b),

$$\lambda_{\max}(\hat{Q}_\tau^{-1}) = O_p(\min\left\{n^{2\delta},\tau_n^{-1}\right\}). \qquad (11.34)$$

Hence $\left\|\hat{\beta}_\tau-\beta\right\| \leq O_p(\min\{n^\delta,\tau_n^{-1/2}\})\left\|\hat{Q}_\tau^{-1/2}\hat{P}'(y-\hat{P}_\#\beta)/n\right\|$. But, by TR,

$$\begin{aligned}&\left\|\hat{Q}_\tau^{-1/2}\hat{P}'(y-\hat{P}_\#\beta)/n\right\| \\ \leq\ &\left\|\hat{Q}_\tau^{-1/2}\hat{P}'(y-h)/n\right\|+\left\|\hat{Q}_\tau^{-1/2}\hat{P}'(h-\hat{P}_\#\beta)/n\right\| \\ =\ &\left\|\hat{Q}_\tau^{-1/2}\hat{P}'(y-h)/n\right\|+\left\|\hat{Q}_\tau^{-1/2}\hat{P}'(h-\hat{P}\beta-n\tau_n\hat{P}(\hat{P}'\hat{P})^{-1}\beta)/n\right\| \\ \leq\ &\left\|\hat{Q}_\tau^{-1/2}\hat{P}'(y-h)/n\right\|+\left\|\hat{Q}_\tau^{-1/2}\hat{P}'(h-\hat{P}\beta)/n\right\|+\left\|\tau_n\hat{Q}_\tau^{-1/2}\beta\right\|.\end{aligned}$$

The first and second terms, note that $c'\hat{P}'\hat{Q}_\tau^{-1}\hat{P}c \leq c'\hat{P}'\hat{Q}^{-1}\hat{P}c$ for any vector $c$, since $(\hat{Q}^{-1}-\hat{Q}_\tau^{-1})$ is p.s.d. Therefore, by (11.31), (11.32), and (11.33) in Lemma 11.7, we have

$\left\|\hat{Q}_\tau^{-1/2}\hat{P}'(y-h)/n\right\| = O_p(\sqrt{K/n})$ and $\left\|\hat{Q}_\tau^{-1/2}\hat{P}'(h-\hat{P}\beta)/n\right\| = O_p(K^{-\alpha}+\Delta_\pi)$. The third term (squared) is

$$\left\|\tau_n \hat{Q}_\tau^{-1/2}\beta\right\|^2 = \tau_n^2\beta'\hat{Q}_\tau^{-1}\beta \le \tau_n^2\lambda_{\max}(\hat{Q}_\tau^{-1})\left\|\beta\right\|^2 \le \tau_n^2\lambda_{\max}(\hat{Q}_\tau^{-1})B \le \tau_n^2 O_p(\min\left\{n^{2\delta}, \tau_n^{-1}\right\}),$$

by $\beta'\beta \le B$ and (11.34) (last ineq.). Therefore, $\left\|\tau_n\hat{Q}_\tau^{-1/2}\beta\right\| = \tau_n O_p(\min\{n^{\delta}, \tau_n^{-1/2}\})$. Consequently, analogous to the proof of Lemma 11.7,

$$\left\|\hat{h}_\tau - h_0\right\|_{L_2} = O_p\left(\min\{n^\delta, \tau_n^{-1/2}\}\left(\sqrt{\frac{K}{n}} + K^{-s/d_x} + \tau_n\cdot\min\{n^\delta,\tau_n^{-1/2}\} + \sqrt{\frac{L}{n}} + L^{-s_1/d_z}\right)\right).$$

This proves the first part of the theorem. The conclusion of the second part follows from

$$\begin{aligned}\sup_w\left|\hat{h}_\tau(w)-h_0(w)\right| &\le \sup_w\left|p^K(w)'\beta - h_0(w)\right| + \sup_w\left|p^K(w)'(\hat{\beta}_\tau-\beta)\right| \\ &\le O(K^{-s}) + \zeta_0^v(K)\left\|\hat{\beta}_\tau - \beta\right\|.\end{aligned}$$

□

**Proof of Theorem 6.3:** The proof follows directly from the proofs of Theorems 4.2 and 4.3 of NPV (p.602). As for notations, we use $v$ instead of $u$ of NPV, and the remaining notations are identical. □

## 11.6 Proof of Asymptotic Normality (Section 7)

Assumption G in Section 7 implies the following technical assumption. Recall that $\zeta_r^v(K)$ satisfies $\sup_{v\in\mathcal{V}}\left\|\partial^r p^K(v)\right\| \le \zeta_r^v(K)$ and, with $\max_{|\mu|\le r}\sup_{w\in\mathcal{W}}\left\|\partial^\mu p^K(w)\right\| \le \zeta_r(K)$, we have $\zeta_r(K)\le\zeta_r^v(K)$.

**Assumption G.1** *The following converge to zero as $n\to\infty$: $\sqrt{n}K^{-\alpha}$, $\sqrt{n}L^{-\alpha_1}$, $\sqrt{n}\tau_n n^\delta$, $\zeta_0^v(K)L/\sqrt{n}$, $n^\delta L^{1/2}\left\{L\zeta_1^v(\kappa) + K^{1/2}\xi(L)\right\}/\sqrt{n}$, $n^{3\delta}K\zeta_1^v(K)L^{1/2}/\sqrt{n}$, $n^{4\delta}\left\{\zeta_0^v(K)^2K + \xi(L)^2L\right\}/n$, and $n^\delta\zeta_0^v(K)L^{1/2}\zeta_1^v(K)(K+L)^{1/2}/\sqrt{n}$.*

**Proof that Assumption G implies Assumption G.1:** By (11.4) and (11.5), $\zeta_r^v(K) = K^{1+2r}$ and $\xi(L) = L$ for power series and $\zeta_r^v(K) = K^{1/2+r}$ and $\xi(L) = L^{1/2}$ for splines. Therefore, for power series

$$\begin{aligned}\zeta_0^v(K)L/\sqrt{n} &= n^{-1/2}KL \\ n^\delta L^{1/2}\left\{L\zeta_1^v(\kappa)+K^{1/2}\xi(L)\right\}/\sqrt{n} &= n^{\delta-\frac{1}{2}}L^{1/2}\left\{LK^3 + K^{1/2}L\right\} \\ &= n^{\delta-\frac{1}{2}}L^{3/2}\left\{K^3+K^{1/2}\right\} \le Cn^{\delta-\frac{1}{2}}K^3L^{3/2} \\ n^{3\delta}K\zeta_1^v(K)L^{1/2}/\sqrt{n} &= n^{3(\delta-\frac{1}{6})}K^4L^{1/2} \\ n^{4\delta}\left\{\zeta_0^v(K)^2K+\xi(L)^2L\right\}/n &= n^{4(\delta-\frac{1}{4})}\left\{K^3+L^3\right\} \\ n^\delta\zeta_0^v(K)L^{1/2}\zeta_1^v(K)(K+L)^{1/2}/\sqrt{n} &\le n^{\delta-\frac{1}{2}}K^4L^{1/2}(K+L)^{1/2},\end{aligned}$$

and for splines

$$
\begin{aligned}
\zeta_0^v(K)L/\sqrt{n} &= n^{-1/2}K^{1/2}L \\
n^{\delta}L^{1/2}\left\{L\zeta_1^v(\kappa)+K^{1/2}\xi(L)\right\}/\sqrt{n} &= n^{\delta-\frac{1}{2}}L\left\{K^{3/2}L^{1/2}+K^{1/2}\right\} \\
&\leq n^{\delta-\frac{1}{2}}K^{3/2}L\left\{L^{1/2}+1\right\}\leq Cn^{\delta-\frac{1}{2}}K^{3/2}L^{3/2} \\
n^{3\delta}K\zeta_1^v(K)L^{1/2}/\sqrt{n} &= n^{3(\delta-\frac{1}{6})}K^{5/2}L^{1/2} \\
n^{4\delta}\left\{\zeta_0^v(K)^2K+\xi(L)^2L\right\}/n &= n^{4(\delta-\frac{1}{4})}\left\{K^2+L^2\right\} \\
n^{\delta}\zeta_0^v(K)L^{1/2}\zeta_1^v(K)(K+L)^{1/2}/\sqrt{n} &\leq n^{\delta-\frac{1}{2}}K^2L^{1/2}(K+L)^{1/2}.
\end{aligned}
$$

Then the results of Assumption G.1 follow. □

**Preliminary derivations for the proof of Theorem 7.1 :** Recall $\Delta_\pi=\sqrt{L/n}+L^{-\alpha_1}$, $\Delta_{\hat{Q}}=\zeta_1^v(\kappa)^2\Delta_\pi^2+\kappa^{1/2}\zeta_1^v(\kappa)\Delta_\pi$, $\Delta_{\tilde{Q}}=\sqrt{\left\{\zeta_1^v(\kappa)^2+\zeta_0^v(\kappa)^2\right\}K/n}$, and $R_n=\min\left\{n^{\delta},\tau_n^{-1/2}\right\}$, and define

$$
\begin{aligned}
\Delta_Q &= \Delta_{\hat{Q}}+\Delta_{\tilde{Q}}, \qquad \Delta_{Q_\tau}=\Delta_Q+\tau_n\sqrt{K}, \qquad \Delta_{Q_1}=\xi(L)L^{1/2}/\sqrt{n} \\
\Delta_H &= L^{1/2}\zeta_1^v(\kappa)\Delta_\pi+K^{1/2}\xi(L)/\sqrt{n}, \qquad \Delta_h=R_n(\sqrt{K/n}+K^{-\alpha}+\Delta_\pi).
\end{aligned}
$$

First, $n^{3\delta}K^{1/2}\Delta_Q\to 0$ implies $n^{2\delta}\Delta_Q\to 0$. Also note that, by $\sqrt{n}K^{-\alpha}\to 0$ and $\sqrt{n}L^{-\alpha_1}\to 0$, we have $\Delta_\pi=(L^{1/2}/\sqrt{n})(1+L^{-1/2}\sqrt{n}L^{-\alpha_1})=O(L^{1/2}/\sqrt{n})$ and $\Delta_h=R_n(\sqrt{K/n}+\sqrt{L/n})$. Therefore, Assumption G.1 implies $\sqrt{n}\zeta_0^v(K)\Delta_\pi^2=O(\zeta_0^v(K)L/\sqrt{n})=o(1)$. Also

$$
\begin{aligned}
n^{\delta}L^{1/2}\Delta_H &= n^{\delta}L^{1/2}\left\{L^{1/2}\zeta_1^v(\kappa)\Delta_\pi+K^{1/2}\xi(L)/\sqrt{n}\right\} \\
&= O(n^{\delta}L^{1/2}\left\{L^{1/2}\zeta_1^v(\kappa)L^{1/2}/\sqrt{n}+K^{1/2}\xi(L)/\sqrt{n}\right\})=o(1) \\
n^{3\delta}\kappa\zeta_1^v(\kappa)\Delta_\pi &= O_p(n^{3\delta}K\zeta_1^v(K)L^{1/2}/\sqrt{n})=o_p(1)
\end{aligned}
$$

by G.1. These results imply $n^{\delta}\zeta_1^v(K)^2\Delta_\pi^2=O(n^{\delta}\zeta_1^v(K)^2L/n)=o(1)$ and also

$$
\begin{aligned}
n^{3\delta}\kappa^{1/2}\Delta_{\hat{Q}} &= n^{3\delta}\kappa^{1/2}\left\{\zeta_1^v(\kappa)^2\Delta_\pi^2+\kappa^{1/2}\zeta_1^v(\kappa)\Delta_\pi\right\}\to 0 \\
n^{3\delta}\kappa^{1/2}\Delta_{\tilde{Q}} &= n^{3\delta}\kappa^{1/2}\sqrt{\left\{\zeta_1^v(\kappa)^2+\zeta_0^v(\kappa)^2\right\}K/n}\leq Cn^{3\delta}\kappa\zeta_1^v(\kappa)/\sqrt{n}\to 0
\end{aligned}
$$

since $n^{3\delta}\kappa^{1/2}\zeta_1^v(\kappa)^2\Delta_\pi^2\to 0$ and $n^{3\delta}\kappa\zeta_1^v(\kappa)\Delta_\pi\to 0$. Also, since $\sqrt{n}\tau_n n^{\delta}\to 0$ and $n^{\delta}K/\sqrt{n}\to 0$ imply $n^{2\delta}K\tau_n=\sqrt{n}\tau_n n^{\delta}\cdot n^{\delta}K/\sqrt{n}\to 0$, consequently, $n^{2\delta}\kappa^{1/2}\Delta_{Q_\tau}\to 0$. Also G.1 assumes $O(n^{4\delta-1}\left\{\zeta_0^v(K)^2K+\xi(L)^2L\right\})=o(1)$. And $L^{1/2}\Delta_{Q_1}\to 0$ is also implied by G.1. Also, since $\sqrt{n}\tau_n n^{\delta}=o(1)$ implies $R_n=n^{\delta}$,

$$
\zeta_0^v(K)L^{1/2}\zeta_1^v(K)\Delta_h=n^{\delta}\zeta_0^v(K)L^{1/2}\zeta_1^v(K)(K+L)^{1/2}/\sqrt{n}\to 0
$$

by G.1, which in turn gives $\zeta_0^v(K)\Delta_h\to 0$. Then with the results above, we also have $n^{2\delta}\left\{\Delta_Q+\zeta_0^v(K)^2K/n\right\}\to 0$. Lastly, $\sqrt{n}\tau_n n^{\delta}\to 0$ implies $n^{2\delta}\tau_n\to 0$ since $n^{\delta}/\sqrt{n}\to 0$, and hence $R_n=n^{\delta}$.

**Proof of Theorem 7.1 :** Given the convergence rate proof above, the proof here is a mild modification of the proof of Theorem 5.1 in NPV (with their trimming function being an identity function). The components established in the convergence rate proof which are distinct from NPV, are used here. The rest of the notations are the same as those of NPV. We prove the theorem under each case of Assumption F(a) and (b). Let 'MVE' abbreviate mean value expansion.

Let $X$ is a vector of variables that includes $x$ and $z$, and $w(X,\pi)$ a vector of functions of $X$ and $\pi$. Note that $w(\cdot,\cdot)$ is a vector of transformation functions of regressors and, trivially, $w(X,\pi)=(x,x-\Pi(z))$. Recall

$$\begin{aligned}\hat{V}_\tau &= A\hat{Q}_\tau^{-1}\left(\hat{\Sigma}+\hat{H}\hat{Q}_1^{-1}\hat{\Sigma}_1\hat{Q}_1^{-1}\hat{H}'\right)\hat{Q}_\tau^{-1}A',\\ \hat{H}_\tau &= \sum \hat{p}_i\left\{\left[\partial\hat{h}_\tau(\hat{w}_i)/\partial w\right]'\partial w(X_i,\hat{\Pi}_i)/\partial\pi\right\}r_i'/n,\end{aligned}$$

and $\hat{Q}=\hat{P}'\hat{P}/n$, $\hat{Q}_\tau=\hat{Q}+\tau_n I$, $Q=E[p_ip_i']$, $\hat{Q}_1=R'R/n$, $\hat{\Sigma}_\tau=\sum\hat{p}_i\hat{p}_i'[y_i-\hat{h}_\tau(\hat{w}_i)]^2/n$, and $\hat{\Sigma}_1=\sum\hat{v}_i^2r_ir_i'/n$, where $r_i=r^L(z_i)$. Then define

$$\begin{aligned}V &= AQ^{-1}\left(\Sigma+HQ_1^{-1}\Sigma_1Q_1^{-1}H'\right)Q^{-1}A',\\ \Sigma &= E[p_ip_i'var(y_i|X_i)],\qquad H=E\left[p_i\left\{[\partial h(w_i)/\partial w]'\partial w(X_i,\Pi_i)/\partial\pi\right\}r_i'\right],\end{aligned}$$

where $V$ does not depend on $\tau$. Note that $H$ is a channel through which the first stage estimation error kicks into the variance of the estimator of $h(\cdot)$ in the outcome equation.

We first prove

$$\sqrt{n}V^{-1/2}(\hat{\theta}_\tau-\theta_0)\to_d N(0,1).$$

For notational simplicity, let $F=V^{-1/2}$. Let $h=(h(w_1),...,h(w_n))'$ and $\tilde{h}=(h(\hat{w}_1),...,h(\hat{w}_n))'$. Also let $\eta_i=y_i-h_0(w_i)$ and $\eta=(\eta_1,...,\eta_n)'$. Let $\Pi=(\Pi_1,...,\Pi_n)'$, $v_i=x_i-\Pi_i$, and $U=(v_1,...,v_n)'$.

As an overview of the proof, note that we prove that $\left\|FAQ^{-1}\right\|=O(n^\delta)$, $\sqrt{n}F[a(p^{K\prime}\tilde{\beta})-a(h_0)]=o_p(1)$, $\sqrt{n}FA(\hat{P}'\hat{P}+n\tau_nI)^{-1}\hat{P}'(\tilde{h}-\hat{P}_{\#}\tilde{\beta})=o_p(1)$, $FA\hat{Q}_\tau^{-1}\hat{P}'(h-\tilde{h})/\sqrt{n}=FAQ^{-1}\times HR'U/\sqrt{n}+o_p(1)$, and $FA\hat{Q}_\tau^{-1}\hat{P}'\eta/\sqrt{n}=FAQ^{-1}\hat{P}'\eta/\sqrt{n}+o_p(1)$ below. If they hold, then we will have, by letting $\tilde{\beta}$ be such that $\left|p^K(\cdot)'\tilde{\beta}-h_0(\cdot)\right|_\delta=O(K^{-\alpha})$,

$$\begin{aligned}\sqrt{n}V^{-1/2}\left(\hat{\theta}_\tau-\theta_0\right) &= \sqrt{n}F\left(a(\hat{h}_\tau)-a(h_0)\right)\\ &= \sqrt{n}F\left(a(p^{K\prime}\hat{\beta}_\tau)-a(p^{K\prime}\tilde{\beta})+a(p^{K\prime}\tilde{\beta})-a(h_0)\right)\\ &= \sqrt{n}FA\hat{\beta}_\tau-\sqrt{n}FA\tilde{\beta}+o_p(1)\\ &= \sqrt{n}FA(\hat{P}'\hat{P}+n\tau_nI)^{-1}\hat{P}'(h+\eta)-\sqrt{n}FA(\hat{P}'\hat{P}+n\tau_nI)^{-1}\hat{P}'\tilde{h}\\ &\quad+\sqrt{n}FA(\hat{P}'\hat{P}+n\tau_nI)^{-1}\hat{P}'(\tilde{h}-\hat{P}_{\#}\tilde{\beta})+o_p(1)\\ &= FA\hat{Q}_\tau^{-1}\hat{P}'\eta/\sqrt{n}-FA\hat{Q}_\tau^{-1}\hat{P}'(h-\tilde{h})/\sqrt{n}+o_p(1)\\ &= FAQ^{-1}(P'\eta/\sqrt{n}+HR'U/\sqrt{n})+o_p(1). \qquad (11.35)\end{aligned}$$

Then, for any vector $\phi$ with $\|\phi\| = 1$, let $Z_{in} = \phi' FAQ^{-1}[p_i\eta_i + Hr_iu_i]/\sqrt{n}$. Note $Z_{in}$ is i.i.d. for each $n$. Also $EZ_{in} = 0$, $var(Z_{in}) = 1/n$ (recall $\Sigma = E[p_ip_i'var(y_i|X_i)]$). Furthermore, $\left\|FAQ^{-1}\right\| = O(n^\delta)$ and $\left\|FAQ^{-1}H\right\| \leq C\left\|FAQ^{-1}\right\| = O(n^\delta)$ by $CI - HH'$ p.s.d, so that, for any $\varepsilon > 0$,

$$\begin{aligned}
nE\left[1\{|Z_{in}| > \varepsilon\} Z_{in}^2\right] &= n\varepsilon^2 E\left[1\{|Z_{in}/\varepsilon| > 1\}(Z_{in}/\varepsilon)^2\right] \\
&\leq n\varepsilon^2 E\left[(Z_{in}/\varepsilon)^4\right] \\
&\leq \frac{n\varepsilon^2}{n^2\varepsilon^4}\|\phi\|^4 \left\|FAQ^{-1}\right\|^4 \{\|p_i\|^2 E\left[\|p_i\|^2 E[\eta_i^4|X_i]\right] \\
&\quad + \|r_i\|^2 E\left[\|r_i\|^2 E[u_i^4|z_i]\right]\} \\
&\leq CO(n^{4\delta})\left\{\zeta_0^v(K)^2 E\|p_i\|^2 + \xi(L)^2 E\|r_i\|^2\right\}/n \\
&\leq CO(n^{4\delta})\left\{\zeta_0^v(K)^2 tr(Q) + \xi(L)^2 tr(Q_1)\right\}/n \\
&\leq O(n^{4\delta-1}\left\{\zeta_0^v(K)^2 K + \xi(L)^2 L\right\}) = o(1)
\end{aligned}$$

by G.1. Then, $\sqrt{n}F(\hat{\theta}_\tau - \theta_0) \to_d N(0,1)$ by Lindbergh-Feller theorem and (11.35).

Now, we proceed with detailed proofs. For simplicity as before, the remainder of the proof will be given for the scalar $\Pi(z)$ case. In the first part, we prove under Assumption F(b) and then F(a). First, suppose Assumption F(b) is satisfied. By CS, $|a(h)| = |A\beta| \leq \|A\|\|\beta\| = \|A\|\left(Eh(x)^2\right)^{1/2}$ so $\|A\| \to \infty$. Since $\lambda_{\min}(Q^{-1})$ is bounded away from zero, $CI = \lambda_{\min}(Q^{-1})I \leq Q^{-1}$. And also since $\sigma^2(X) = var(y|X)$ is bounded away from zero by Assumption E, we have $\Sigma \geq CQ$. Hence

$$V \geq AQ^{-1}\Sigma Q^{-1}A' \geq CAQ^{-1}QQ^{-1}A' \geq CAQ^{-1}A' \geq \tilde{C}\|A\|^2, \tag{11.36}$$

Therefore, it follows that $F$ is bounded.

Now, instead, suppose Assumption F(a) is satisfied. Then $\underset{1\times K}{A} = a(p^K) = E[\nu(x_i)p^K(w_i)']$. Let $\nu_K = \nu_K(w) = AQ^{-1}p^K(w) = E[\nu(x_i)p^K(w_i)']E[p^K(w_i)p^K(w_i)']^{-1}p^K(w_i)$, which is (transpose of) mean square projection of $\nu(\cdot)$ on approximating functions $(p^K(\cdot))$. Then, $E\|\nu - \nu_K\|^2 \leq E\left\|\nu - \beta_K' p^K\right\|^2 \to 0$. Let $d(X) = [\partial h(w_i)/\partial w]'\partial w(X_i, \Pi_i)/\partial\pi$, $b_{KL}(z) = E[d(X)\nu_K(w)r^L(z)']r^L(z)$ and $b_L(z) = E[d(X)\nu(w)r^L(z)']r^L(z)$. Then

$$E[\|b_{KL}(z) - b_L(z)\|^2] \leq E[d(X)^2\|\nu_K(w) - \nu(w)\|^2] \leq CE[\|\nu_K(w) - \nu(w)\|^2] \to 0$$

as $K \to \infty$. Furthermore by Assumption E, $E[\|b_L(z) - \rho(z)\|^2] \to 0$ as $L \to \infty$, where $\rho(z)$ is a matrix of projections of elements of $\nu(w)d(X)$ on $\mathcal{L}$ which is the set of limit points of $r^L(z)'\gamma_L$. Then (as in (A.10) of NPV), by Assumption E

$$V = E[\nu_K(w)\nu_K(w)'\sigma^2(X)] + E[b_{KL}(z)var(x|z)b_{KL}(z)'] \to \bar{V},$$

where $\bar{V} = E[\nu(w)\nu(w)'\sigma^2(X)] + E[\rho(z)var(x|z)\rho(z)']$. This shows $F$ is bounded.

Next, by the previous proofs on the convergence rate,

$$\begin{aligned}\left\|\hat{Q}_\tau - Q\right\| &\le \left\|\hat{Q} - Q + \tau_n I\right\| \le \left\|\hat{Q} - \tilde{Q}\right\| + \left\|\tilde{Q} - Q\right\| + \|\tau_n I\| \\ &\le O_p(\Delta_{\hat{Q}}) + O_p(\Delta_{\tilde{Q}}) + O(\tau_n \sqrt{K}) = O_p(\Delta_{Q_\tau})\end{aligned}$$

and, by letting $Q_1 = I$, $\left\|\hat{Q}_1 - I\right\| = O_p(\xi(L)L^{1/2}/\sqrt{n}) = O_p(\Delta_{Q_1})$. Furthermore, with $\bar{H} = \sum \hat{p}_i d(X_i) r_i'/n$, similarly to the proofs above $\left\|\bar{H} - H\right\| = O_p(\Delta_H) = o_p(1)$, where $\Delta_H = \left[L^{1/2}\zeta_1(K) + \zeta_0(K)\xi(L)^2\right]\Delta_\pi + K^{1/2}\xi(L)/\sqrt{n}$ as in NPV. Now, by (11.36)

$$\left\|FAQ^{-1/2}\right\|^2 = tr(FAQ^{-1}A'F) \le tr(CFVF) = C.$$

By Assumption G.1, $R_n = n^\delta$ and hence by (11.34) and Lemma 11.7(a), $\lambda_{\max}(\hat{Q}_\tau^{-1}) = O_p(n^{2\delta})$ and $\lambda_{\max}(Q^{-1}) = O(n^{2\delta})$, respectively. And then,

$$\left\|FAQ^{-1}\right\| = \left\|FAQ^{-1/2}Q^{-1/2}\right\| \le \lambda_{\max}(Q^{-1})^{1/2}\left\|FAQ^{-1/2}\right\| \le CO(n^\delta).$$

Note that for any matrix $B$, $\left\|B\hat{Q}_\tau^{-1}\right\| \le \left\|B\hat{Q}^{-1}\right\|$, since $(\hat{Q}^{-1} - \hat{Q}_\tau^{-1})$ is p.s.d. Hence

$$\begin{aligned}\left\|FA'\hat{Q}_\tau^{-1}\right\| &\le \left\|FA'\hat{Q}^{-1}\right\| \le \left\|FAQ^{-1}\right\| + \left\|FAQ^{-1}\left(\hat{Q} - Q\right)\hat{Q}^{-1}\right\| \\ &\le CO(n^\delta) + CO(n^\delta)O_p(n^{2\delta})\left\|\hat{Q} - Q\right\| \\ &= O(n^\delta) + O_p(n^{3\delta}\Delta_Q) = O(n^\delta) + o_p(1) = O_p(n^\delta)\end{aligned}$$

by Assumption G.1. Also

$$\begin{aligned}\left\|FA\hat{Q}_\tau^{-1/2}\right\|^2 &\le \left\|FA'\hat{Q}^{-1/2}\right\|^2 \le \left\|FAQ^{-1/2}\right\|^2 + tr(FAQ^{-1}\left(\hat{Q} - Q\right)\hat{Q}^{-1}A'F) \\ &\le C + \left\|FA'Q^{-1}\right\|\left\|\hat{Q} - Q\right\|\left\|FA'\hat{Q}^{-1}\right\| \\ &\le O(n^\delta)O_p(\Delta_Q)O(n^\delta) = o_p(1).\end{aligned}$$

Firstly, as $\tilde{\beta}$ is defined such that $\left|p^K(\cdot)'\tilde{\beta} - h_0(\cdot)\right|_\delta = O(K^{-\alpha})$,

$$\begin{aligned}\left\|\sqrt{n}F\left[a(p^{K\prime}\tilde{\beta}) - a(h_0)\right]\right\| &= \left\|\sqrt{n}F\left[a(p^{K\prime}\tilde{\beta} - h_0)\right]\right\| \le \sqrt{n}\,|F|\left|p^K(\cdot)'\tilde{\beta} - h_0(\cdot)\right|_\delta \\ &\le C\sqrt{n}K^{-\alpha} = o_p(1)\end{aligned}$$

by G.1. Secondly,

$$\begin{aligned}\left\|FA\hat{Q}_\tau^{-1}\hat{P}'(\tilde{h} - \hat{P}_\#\tilde{\beta})/\sqrt{n}\right\| &\le \left\|FA\hat{Q}_\tau^{-1}\hat{P}'/\sqrt{n}\right\|\sqrt{n}\sup_W\left|p^K(w)'\tilde{\beta} - h_0(w)\right| \\ &\quad + \left\|n\tau_n FA\hat{Q}_\tau^{-1}\beta/\sqrt{n}\right\| \\ &\le \left\|FA\hat{Q}_\tau^{-1/2}\right\|\sqrt{n}O(K^{-\alpha}) + \sqrt{n}\tau_n\left\|FA\hat{Q}_\tau^{-1}\right\|\|\beta\| \\ &\le o_p(1)O(\sqrt{n}K^{-\alpha}) + O_p(\sqrt{n}\tau_n n^\delta) = o_p(1)\end{aligned}$$

by G.1. Thirdly, let $\gamma$ be such that $\sup_Z \left|\Pi_0(z) - r^L(z)'\gamma\right| = O(L^{-\alpha_1})$, and $d_i = d(X_i)$. By a second order MVE of each $h(\hat{w}_i)$ around $w_i$

$$\begin{aligned} FA\hat{Q}_\tau^{-1}\hat{P}'(h-\tilde{h})/\sqrt{n} &= FA\hat{Q}_\tau^{-1}\sum_i \hat{p}_i d_i[\hat{\Pi}_i - \Pi_i]/\sqrt{n} + \hat{\rho} \\ &= FA\hat{Q}_\tau^{-1}\bar{H}\hat{Q}_1^{-1}R'U/\sqrt{n} + FA\hat{Q}_\tau^{-1}\bar{H}\hat{Q}_1^{-1}R'(\Pi - R'\gamma)/\sqrt{n} \\ &\quad + FA\hat{Q}_\tau^{-1}\sum_i \hat{p}_i d_i[r_i'\gamma - \Pi_i]/\sqrt{n} + \hat{\rho}. \end{aligned}$$

But $\|\hat{\rho}\| \leq C\sqrt{n}\left\|FA\hat{Q}_\tau^{-1/2}\right\|\zeta_0^v(K)\sum_i \left\|\hat{\Pi}_i - \Pi_i\right\|^2/n = o_p(1)O_p(\sqrt{n}\zeta_0^v(K)\Delta_\pi^2) = o_p(1)$. Also, by $d_i$ being bounded and $n\bar{H}\hat{Q}_1^{-1}\bar{H}'$ being equal to the matrix sum of squares from the multivariate regression of $\hat{p}_i d_i$ on $r_i$, $\bar{H}\hat{Q}_1^{-1}\bar{H}' \leq \sum_i \hat{p}_i\hat{p}_i' d_i^2/n \leq C\hat{Q} \leq C\hat{Q}_\tau$. Therefore,

$$\begin{aligned} & \left\|FA\hat{Q}_\tau^{-1}\bar{H}\hat{Q}_1^{-1}R'(\Pi - R'\gamma)/\sqrt{n}\right\| \\ \leq & \left\|FA\hat{Q}_\tau^{-1}\bar{H}\hat{Q}_1^{-1}R'/\sqrt{n}\right\|\sqrt{n}\sup_Z\left|\Pi_0(z) - r^L(z)'\gamma\right| \\ \leq & \left[tr\left(FA\hat{Q}_\tau^{-1}\bar{H}\hat{Q}_1^{-1}\hat{Q}_1\hat{Q}_1^{-1}\bar{H}'\hat{Q}_\tau^{-1}A'F'\right)\right]^{1/2} O(\sqrt{n}L^{-\alpha_1}) \\ \leq & \; C\left[tr\left(FA\hat{Q}_\tau^{-1}\hat{Q}_\tau\hat{Q}_\tau^{-1}A'F'\right)\right]^{1/2} O(\sqrt{n}L^{-\alpha_1}) \\ \leq & \; C\left\|FA\hat{Q}_\tau^{-1/2}\right\| O(\sqrt{n}L^{-\alpha_1}) = o_p(1)O_p(\sqrt{n}L^{-\alpha_1}) = o_p(1) \end{aligned}$$

by G.1. Similarly,

$$\left\|FA\hat{Q}_\tau^{-1}\sum_i \hat{p}_i d_i[r_i'\gamma - \Pi_i]/\sqrt{n}\right\| \leq C\left\|FA\hat{Q}_\tau^{-1/2}\right\| O(\sqrt{n}L^{-\alpha_1}) = o_p(1)O_p(\sqrt{n}L^{-\alpha_1}) = o_p(1).$$

Next, we consider the term $FA\hat{Q}_\tau^{-1}\bar{H}\hat{Q}_1^{-1}R'U/\sqrt{n}$. Note that $E\left\|R'U/\sqrt{n}\right\|^2 = tr(\Sigma_1) \leq Ctr(I_L) \leq L$ by $E[u^2|z]$ bounded, so by MR, $\|R'U/\sqrt{n}\| = O_p(L^{1/2})$. Also, we have

$$\left\|FA\hat{Q}_\tau^{-1}\bar{H}\hat{Q}_1^{-1}\right\| \leq O_p(1)\left\|FA\hat{Q}_\tau^{-1/2}\right\| = o_p(1).$$

Therefore

$$\begin{aligned} \left\|FA\hat{Q}_\tau^{-1}\bar{H}(\hat{Q}_1^{-1} - I)R'U/\sqrt{n}\right\| &\leq \left\|FA\hat{Q}_\tau^{-1}\bar{H}\hat{Q}_1^{-1}\right\|\left\|\hat{Q}_1 - I\right\|\left\|R'U/\sqrt{n}\right\| \\ &= o_p(1)O_p(\Delta_{Q_1})O_p(L^{1/2}) = o_p(1) \end{aligned}$$

by G.1. Similarly,

$$\begin{aligned} \left\|FA\hat{Q}_\tau^{-1}(\bar{H} - H)R'U/\sqrt{n}\right\| &\leq \left\|FA\hat{Q}_\tau^{-1}\right\|\left\|\bar{H} - H\right\|\left\|R'U/\sqrt{n}\right\| \\ &= O_p(n^\delta)O_p(\Delta_H)O_p(L^{1/2}) = o_p(1) \end{aligned}$$

by G.1. Note that $HH'$ is the population matrix mean-square of the regression of $p_i d_i$ on $r_i$ so that $HH' \leq C$, it follows that $E\left\|HR'U/\sqrt{n}\right\|^2 = tr(H\Sigma_1 H') \leq CK$ and therefore,

$\|HR'U/\sqrt{n}\| = O_p(K^{1/2})$. And then,

$$\begin{aligned} \left\|FA(\hat{Q}_\tau^{-1} - Q^{-1})HR'U/\sqrt{n}\right\| &\leq \lambda_{\max}(\hat{Q}_\tau^{-1}) \left\|FAQ^{-1}\right\| \left\|Q - \hat{Q}_\tau\right\| \left\|HR'U/\sqrt{n}\right\| \\ &= O(n^{2\delta})O(n^{\delta})O_p(\Delta_{Q_\tau})O_p(K^{1/2}) = o_p(1). \end{aligned}$$

Combining the results above and by TR, $FA\hat{Q}_\tau^{-1}\hat{P}'(h - \tilde{h})/\sqrt{n} = FAQ^{-1}HR'U/\sqrt{n} + o_p(1)$.

Lastly, similar to (11.28) in the convergence rate part,

$$\left\|\hat{Q}_\tau^{-1/2}(P - \hat{P})'\eta/\sqrt{n}\right\| = O_p(n^{\delta}\zeta_1^{v}(K)^2\Delta_\pi^2) = o_p(1)$$

by G.1 (and by (A.6) of NPV), which implies

$$\left\|FA\hat{Q}_\tau^{-1}(\hat{P} - P)'\eta/\sqrt{n}\right\| \leq \left\|FA\hat{Q}_\tau^{-1/2}\right\| \left\|\hat{Q}_\tau^{-1/2}(P - \hat{P})'\eta/\sqrt{n}\right\| = o_p(1)o_p(1) = o_p(1).$$

Also, by $E[\eta|X] = 0$,

$$\begin{aligned} & E\left[\left\|FA(\hat{Q}_\tau^{-1} - Q^{-1})P'\eta/\sqrt{n}\right\|^2 |X_n\right] \\ \leq\ & tr\left(FA(\hat{Q}_\tau^{-1} - Q^{-1})\left[\sum p_i p_i' var(y_i|X_i)/n\right](\hat{Q}_\tau^{-1} - Q^{-1})A'F\right) \\ \leq\ & Ctr\left(FA(\hat{Q}_\tau^{-1} - Q^{-1})\hat{Q}_\tau(\hat{Q}_\tau^{-1} - Q^{-1})A'F\right) \\ =\ & Ctr\left(FAQ^{-1}(\hat{Q}_\tau - Q)\hat{Q}_\tau^{-1}(\hat{Q}_\tau - Q)Q^{-1}A'F\right) \\ \leq\ & O_p(n^{2\delta})\left\|FAQ^{-1}\right\|^2 \left\|\hat{Q}_\tau - Q\right\|^2 \\ \leq\ & O_p(n^{2\delta}\Delta_{Q_\tau})^2 = o_p(1) \end{aligned}$$

by G.1. Combining all of the previous results and by TR,

$$\sqrt{n}F\left[a(\hat{h}_\tau) - a(h_0)\right] = FAQ^{-1}(P'\eta/\sqrt{n} + HR'U/\sqrt{n}) + o_p(1).$$

Next, recall $V = AQ^{-1}\left(\Sigma + HQ_1^{-1}\Sigma_1 Q_1^{-1}H'\right)Q^{-1}A'$. Note $CI - H\Sigma_1 H'$ is p.s.d (that is $H\Sigma_1 H' \leq CI$ using $Q_1 = I$) and since $var(y|X)$ is bounded by Assumption E, $\Sigma \leq CQ$. Therefore,

$$\begin{aligned} V &= AQ^{-1}\left(\Sigma + HQ_1^{-1}\Sigma_1 Q_1^{-1}H'\right)Q^{-1}A' = AQ^{-1}\left(\Sigma + H\Sigma_1 H'\right)Q^{-1}A' \\ &\leq AQ^{-1}\left(CQ + CI\right)Q^{-1}A' = C\left\|AQ^{-1/2}\right\|^2 + C\left\|AQ^{-1}\right\|^2. \end{aligned}$$

Note that it is reasonable that the first stage estimation error is not cancelled out with $Q^{-1}$, since the first stage regressors does not suffer multicollinearity. Recall $a(p^{K\prime}\beta) = A\beta$ so $a(p^{K\prime}A') = AA'$ ($A$ is row vector). And $\|A\|^2 \leq \left|a(p^{K\prime}A')\right| \leq \left|Ap^K\right|_r \leq \zeta_r(K)\|A\|$ so $\|A\| \leq \zeta_r(K)$. Hence $\left\|AQ^{-1/2}\right\|^2 \leq O(n^{2\delta})\|A\|^2 = O(n^{\delta}\zeta_r(K))^2$ by CS. Thus, $V =$

$O(n^{\delta}\zeta_r(K))^2 + O(n^{2\delta}\zeta_r(K))^2$. Therefore,

$$\hat{\theta}_\tau - \theta_0 = O_p(V^{1/2}/\sqrt{n}) = O_p(n^{2\delta}\zeta_r(K)/\sqrt{n}) = O_p(n^{2(\delta-\frac{1}{4})}\zeta_r(K)) \leq O_p(n^{2(\delta-\frac{1}{4})}\zeta_r^v(K)).$$

This result for scalar $a(h)$ covers the case of Assumption F(b). In the case of F(a), it follows from $V \to \bar{V}$ that $\hat{\theta}_\tau - \theta_0 = O_p(V^{1/2}/\sqrt{n}) = O_p(1/\sqrt{n})$.

Now we can prove

$$\sqrt{n}\hat{V}_\tau^{-1/2}(\hat{\theta}_\tau - \theta_0) \to_d N(0,1)$$

by showing $\left|F\hat{V}_\tau F - 1\right| \to_p 0$. Then, in the case of F(b), $V^{-1}\hat{V}_\tau \to_p 1$, so that $\sqrt{n}\hat{V}_\tau^{-1/2}(\hat{\theta}_\tau - \theta_0) = \sqrt{n}V^{-1/2}(\hat{\theta}-\theta_0)/(V^{-1}\hat{V}_\tau)^{1/2} \to_d N(0,1)$. In other case, the conclusion follows similarly from $F \to \bar{V}^{1/2}$.

For the rest part of the proof can directly be followed by the relevant part of the proof of NPV (p.600-601), except that in our case $Q \neq I$ due to weak instrument. Therefore the following replaces the corresponding part in the proof: For any matrix $B$, we have $\|B\Sigma\| \leq C\|BQ\|$ by $\Sigma \leq CQ$. Therefore,

$$\begin{aligned}
& \left\|FA(\hat{Q}_\tau^{-1}\Sigma\hat{Q}_\tau^{-1} - Q^{-1}\Sigma Q^{-1})A'F'\right\| \\
\leq\ & \left\|FA(\hat{Q}_\tau^{-1} - Q^{-1})\Sigma\hat{Q}_\tau^{-1}A'F'\right\| + \left\|FAQ^{-1}\Sigma(\hat{Q}_\tau^{-1} - Q^{-1})A'F'\right\| \\
\leq\ & \left\|FA\hat{Q}_\tau^{-1}(Q - \hat{Q}_\tau)Q^{-1}\Sigma\hat{Q}_\tau^{-1}A'F'\right\| + \left\|FAQ^{-1}\Sigma Q^{-1}(Q - \hat{Q}_\tau)\hat{Q}_\tau^{-1}A'F'\right\| \\
\leq\ & \left\|FA\hat{Q}_\tau^{-1}\right\|^2 \left\|(Q - \hat{Q}_\tau)Q^{-1}\Sigma\right\| + \left\|FAQ^{-1}\right\| \left\|\Sigma Q^{-1}(Q - \hat{Q}_\tau)\right\| \left\|FA\hat{Q}_\tau^{-1}\right\| \\
\leq\ & C\left\|FA\hat{Q}_\tau^{-1}\right\|^2 \left\|(Q - \hat{Q}_\tau)Q^{-1}Q\right\| + C\left\|FAQ^{-1}\right\| \left\|QQ^{-1}(Q - \hat{Q}_\tau)\right\| \left\|FA\hat{Q}_\tau^{-1}\right\| \\
\leq\ & O_p(n^{2\delta})O_p(\Delta_{Q_\tau}) + O_p(n^{2\delta})O_p(\Delta_{Q_\tau}) = o_p(1)
\end{aligned}$$

by Assumption G.1. Also note that in our proof, $Q_\tau$ is introduced by penalization but the treatment is the same as above. Specifically, one can apply $\left\|B\hat{Q}_\tau^{-1}C\hat{Q}_\tau^{-1}B\right\| \leq \left\|B\hat{Q}^{-1}C\hat{Q}^{-1}B\right\|$ for any matrix $B$ and $C$ of corresponding orders. Also, recall $\zeta_r(K) \leq \zeta_r^v(K)$ and $\Delta_h$ and $\Delta_Q$ are redefined in this paper. That is, by $\zeta_0^v(K)\Delta_h$, $n^{2\delta}\left\{\Delta_Q + \zeta_0^v(K)^2K/n\right\}$, and $\zeta_0^v(K)L^{1/2}\zeta_1^v(K)\Delta_h$ converging to zero, we can prove the following:

$$\left\|FA\hat{Q}_\tau^{-1}(\hat{\Sigma} - \tilde{\Sigma})\hat{Q}_\tau^{-1}A'F'\right\| \leq Ctr(\hat{D})\max_{i\leq n}\left|\hat{h}_i - h_i\right| \leq O_p(1)O_p(\zeta_0^v(K)\Delta_h) = o_p(1)$$

$$\left\|FA\hat{Q}_\tau^{-1}(\tilde{\Sigma} - \Sigma)\hat{Q}_\tau^{-1}A'F'\right\| \leq \left\|FA\hat{Q}_\tau^{-1}\right\|^2 \left\|\tilde{\Sigma} - \Sigma\right\| \leq O_p(n^{2\delta})O_p(\Delta_Q + \zeta_0^v(K)^2K/n) = o_p(1)$$

$$\begin{aligned}
\left\|\hat{H} - \bar{H}\right\| &\leq C\left(\sum_{i=1}^n \|\hat{p}_i\|^2 \|r_i\|^2/n\right)^{1/2} \left(\sum_{i=1}^n \left|\hat{d}_i - d_i\right|^2/n\right)^{1/2} \\
&= O_p(\zeta_0^v(K)L^{1/2})O_p(\zeta_1^v(K)\Delta_h) = o_p(1)
\end{aligned}$$

The rest of the proof thus follows. □

# References

[1] Ai, C., and Chen, X. (2007): "Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models with Different Conditioning Variables," Journal of Econometrics, 141(1), 5-43.

[2] Amemiya, T. (1977): "The ML and the Nonlinear Three-Stage Least Squares Estimator in the General Nonlinear Simultaneous Equations Model," Econometrica, 45, 955–968.

[3] Andrews, D. W. K. (1991): "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models," Econometrica, 59(2), 307-345.

[4] Andrews, D. W. K., and Cheng, X. (2010): "Estimation and Inference with Weak Identification," Cowles Foundation Discussion Paper No.1773, Yale University.

[5] Andrews, D. W. K., and Stock, J. H. (2007): "Inference with Weak Instruments," in Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society, Vol. III, ed. by R. Blundell, W. K. Newey, and T. Persson. Cambridge, UK: Cambridge University Press.

[6] Andrews, D. W. K., and Whang, Y.-J. (2009): "Additive Interactive Regression Models: Circumvention of the Curse of Dimensionality," Econometric Theory, 6(04), 466-479.

[7] Angrist, J. D., and Lavy, V. (1999): "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," The Quarterly Journal of Economics, 114(2), 533-575.

[8] Arlot, S., and Celisse, A. (2010): "A Survey of Cross-Validation Procedures for Model Selection," Statistics Surveys, 4, 40-79.

[9] Blundell, R., Chen, X., and Kristensen, D. (2007): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," Econometrica, 75(6), 1613-1669.

[10] Blundell, R., and Duncan, A. (1998): "Kernel Regression in Empirical Microeconomics," The Journal of Human Resources, 33(1), 62-87.

[11] Blundell, R., Duncan, A., and Pendakur, K. (1998): "Semiparametric Estimation and Consumer Demand," Journal of Applied Econometrics, 13(5), 435-461.

[12] Blundell, R. W., and Powell, J. L. (2004): "Endogeneity in Semiparametric Binary Response Models," Review of Economic Studies, 71(3), 655–679.

[13] Blundell, R., and Powell, J. L. (2006): "Endogeneity in Nonparametric and Semiparametric Regression Models," Cambridge University Press.

[14] Bound, J., Jaeger, D. A., and Baker, R. M. (1995): "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak," Journal of the American statistical association, 90(430), 443-450.

[15] Chen, X, and Pouzo, D. (2009): "Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals," Cowles Foundation Discussion Paper No.1650R, Yale University.

[16] Chernozhukov, V. and Hansen, C. (2005): "An IV Model of Quantile Treatment Effects," Econometrica, 73(1), 245–261.

[17] Chesher, A. (2003): "Identification in Nonseparable Models," Econometrica, 71(5), 1405-1441.

[18] Chesher, A. (2005): "Nonparametric Identification under Discrete Variation," Econometrica, 73(5), 1525-1550.

[19] Das, M., Newey, W. K. and Vella, F. (2003): "Nonparametric Estimation of Sample Selection Models," Review of Economic Studies, 70(1), 33–58.

[20] Dufour, J. M. (1997): "Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models," Econometrica, 65(6), 1365-1387.

[21] Dustmann, C., and Meghir, C. (2005): "Wages, Experience and Seniority," Review of Economic Studies, 72(1), 77-108.

[22] Folland, G. B. (1999): "Real Analysis: Modern Techniques and Their Applications," Wiley.

[23] Garg, K. M. (1998): "Theory of Differentiation," John Wiley & Sons.

[24] Giorgi, G., Guerraggio, A., Thierfelder, J., Thierfelder, J. (2004): "Mathematics of Optimization: Smooth and Nonsmooth Case," Elsevier.

[25] Hall, P., and Horowitz, J. L. (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," The Annals of Statistics, 33(6), 2904-2929.

[26] Horowitz, J. L. and Lee, S. (2007): "Nonparametric Instrumental Variables Estimation of a Quantile Regression Model," Econometrica, 75(4), 1191–1208.

[27] Han, S. (2011): "Identification and Inference in Bivariate Probit Models with Weak Instruments," unpublished manuscript, Department of Economics, Yale University.

[28] Han, C. and Phillips, P. C. B. (2006): "GMM with Many Moment Conditions," Econometrica, 74(1), 147-192.

[29] Hastie, T. J., and Tibshirani, R. J. (1990): "Generalized Additive Models," CRC Press.

[30] Heckman, J. J. (1979): "Sample Selection Bias as a Specification Error," Econometrica, 47(1), 153-161.

[31] Hoerl, A. E., and Kennard, R. W. (1970): "Ridge Regression: Biased Estimation for Nonorthogonal Problems," Technometrics, 12(1), 55-67.

[32] Hong, Y., and White, H. (1995): "Consistent Specification Testing Via Nonparametric Series Regression," Econometrica, 63(5), 1133-1159.

[33] Horowitz, J. L. (2011): "Applied Nonparametric Instrumental Variables Estimation," Econometrica, 79(2), 347-394.

[34] Imbens, G. W., and Newey, W. K. (2009): “Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity,” Econometrica, 77(5), 1481-1512.

[35] Jiang, J., Fan, Y., and Fan, J. (2010): “Estimation in Additive Models with Highly or Nonhighly Correlated Covariates,” The Annals of Statistics, 38(3), 1403-1432.

[36] Jun, S. J., and Pinkse, J. (2007): “Weak Identification and Conditional Moment Restrictions,” The Pennsylvania State University.

[37] Kleibergen, F. (2002): “Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression,” Econometrica, 70(5), 1781–1803.

[38] Kleibergen, F. (2005): “Testing Parameters in GMM Without Assuming that They Are Identified,” Econometrica, 73(4), 1103–1123.

[39] Kress, R. (1989): “Linear Integral Equations,” Springer-Verlag, New York.

[40] Lee, J. M. (2011): “Introduction to Topological Manifolds,” Springer.

[41] Lee, S. (2007): “Endogeneity in Quantile Regression Models: A Control Function Approach,” Journal of Econometrics, 141(2), 1131-1158.

[42] Lorentz, G. G. (1986): “Bernstein Polynomials,” Chelsea, New York.

[43] Newey, W. K. (1990): “Efficient Instrumental Variables Estimation of Nonlinear Models,” Econometrica, 58(4), 809-837.

[44] Newey, W. K. (1997): “Convergence Rates and Asymptotic Normality for Series Estimators,” Journal of Econometrics, 79(1), 147-168.

[45] Newey, W. K., and Powell, J. L. (2003): “Instrumental Variable Estimation of Nonparametric Models,” Econometrica, 71(5), 1565-1578.

[46] Newey, W. K., Powell, J. L., and Vella, F. (1999): “Nonparametric Estimation of Triangular Simultaneous Equations Models,” Econometrica, 67(3), 565-603.

[47] Newey, W. K., and Windmeijer, F. (2009): “Generalized Method of Moments With Many Weak Moment Conditions,” Econometrica 77(3), 687-719.

[48] Powell, M. J. D. (1981): “Approximation Theory and Methods,” Cambridge University Press, Cambridge, UK.

[49] Rivers, D., and Vuong, Q. H. (1988): “Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models,” Journal of Econometrics, 39(3), 347-366.

[50] Schumaker, L. L. (1981): “Spline Functions: Basic Theory,” Wiley, New York.

[51] Smith, R. J., and Blundell, R. W. (1986): “An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply,” Econometrica, 54(3), 679-685.

[52] Staiger, D., and Stock, J. H. (1997): “Instrumental Variables Regression with Weak Instruments,” Econometrica, 65(3), 557-586.

[53] Stock, J. H., and Wright, J. H. (2000): "Instrumental Variables Regression with Weak Instruments," Econometrica, 68(5), 1055-1096.

[54] Stock, J. H., and Yogo, M. (2005): "Testing for Weak Instruments in Linear IV Regression," in Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg, ed. by D. W. K. Andrews and J. H. Stock. Cambridge, U.K.: Cambridge University Press, 80–108.

[55] Stone, C. J. (1982): "Optimal Global Rates of Convergence for Nonparametric Regression," The Annals of Statistics, 10(4), 1040-1053.

[56] Taylor, A. E. (1985): "General Theory of Functions and Integration," Courier Dover Publications.

[57] Weyl, H. (1912): "Das asymtotische Verteilungsgesetz der Eigenwerte lineare partieller Differentialgleichungen," Math. Ann. 71, 441-479.

< SUMMARY >

Figures 4-5: PIV versus IV estimators ($\hat{g}_\tau(\cdot)$ vs. $\hat{g}(\cdot)$)

Figures 6-7: PIV versus IV estimators ($\hat{g}_\tau(\cdot)$ vs. $\hat{g}(\cdot)$), different $\tau$

Figures 8-9: Distributions of $\hat{\theta} = a(\hat{g})$

Table 2: Integrated squared bias, integrated variance and integrated MSE of $\hat{g}_\tau(\cdot)$ and $\hat{g}(\cdot)$

Figure 10: Maimonides' rule

Figures 11-12: Control function versus NPIV estimates of class-size effect

Table 3: Cross-validation criterion

Figure 13: PIV estimates of class-size effect

Figures 14-15: Linear versus nonparametric reduced form for IV estimates of class-size effect

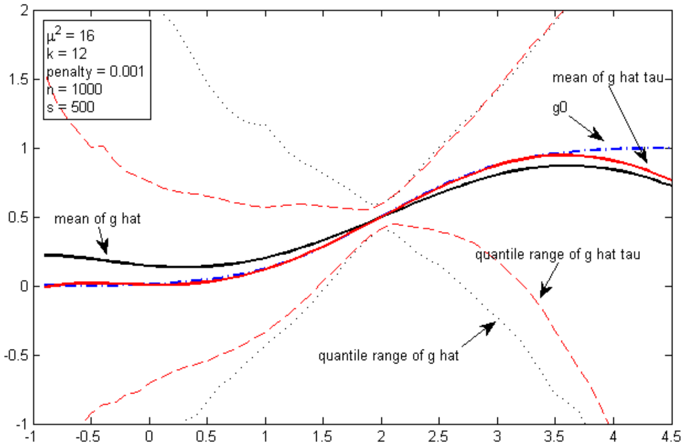

Figure 4: Penalized versus unpenalized estimators ($\hat{g}_\tau(\cdot)$ vs. $\hat{g}(\cdot)$) with a weak instrument, $\tau = 0.001$.

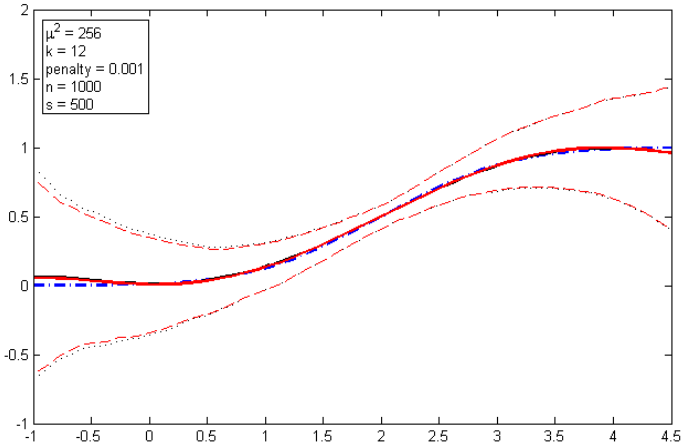

Figure 5: Penalized versus unpenalized estimators ($\hat{g}_\tau(\cdot)$ vs. $\hat{g}(\cdot)$) with a strong instrument, $\tau = 0.001$.

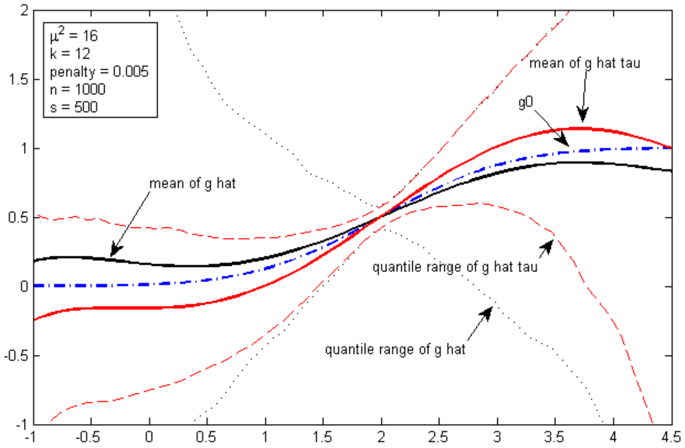

Figure 6: Penalized versus unpenalized estimators ($\hat{g}_\tau(\cdot)$ vs. $\hat{g}(\cdot)$) with a weak instrument, $\tau = 0.005$.

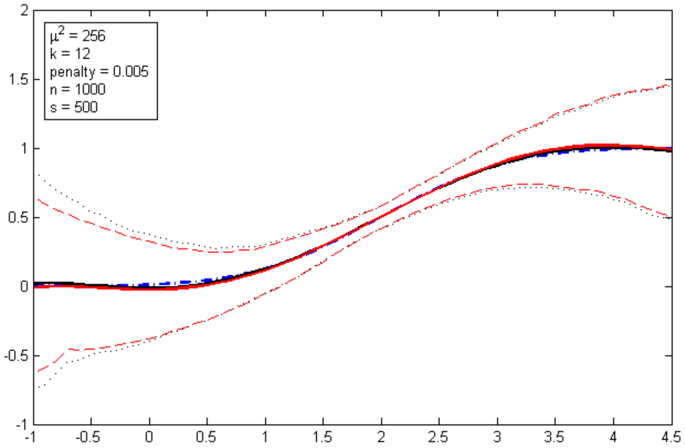

Figure 7: Penalized versus unpenalized estimators ($\hat{g}_\tau(\cdot)$ vs. $\hat{g}(\cdot)$) with a strong instrument, $\tau = 0.005$.

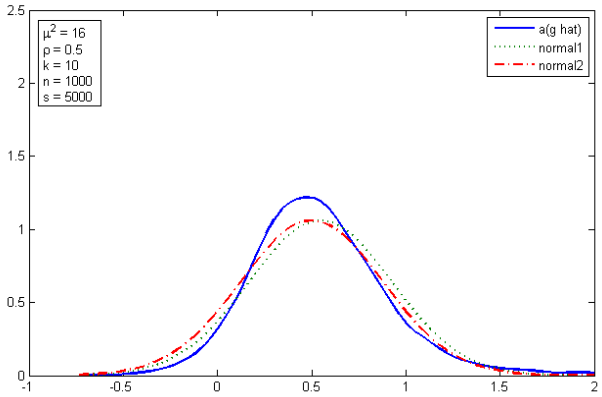

Figure 8: Distribution of $\hat{\theta}$ when the instrument is “nonparametrically” weak.

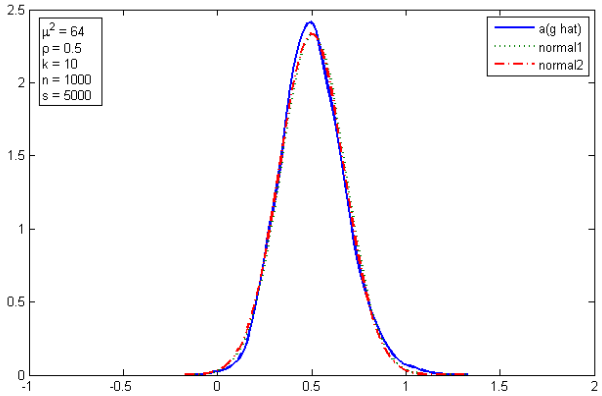

Figure 9: Distribution of $\hat{\theta}$ when the instrument is “nonparametrically” strong.

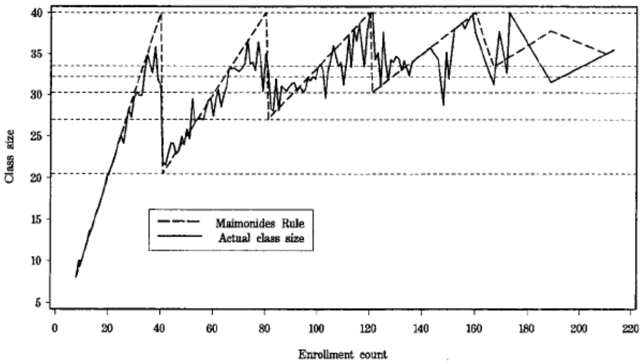

Figure 10: Class size by enrollment count (Angrist and Lavy (1999)).

| $\rho = 0.5$ | | $\mu^2$ | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 4 | 8 | 16 | 32 | 64 | 128 | 256 |
| | $Bias^2$ | 0.0343 | 0.0463 | 0.0032 | 0.0006 | 0.0002 | 0.0001 | 0.0000 |
| IV | $Var$ | 830.0765 | 20.7368 | 0.3220 | 0.0971 | 0.0450 | 0.0273 | 0.0188 |
| | $MSE$ | 830.1108 | 20.7831 | 0.3252 | 0.0977 | 0.0451 | 0.0274 | 0.0188 |
| $Bias^2_{IV}/Bias^2_{LS}$ | | 0.1399 | 0.1733 | 0.0138 | 0.0025 | 0.0007 | 0.0005 | 0.0000 |
| $MSE_{IV}/MSE_{LS}$ | | 3138.0 | 75.6541 | 1.3755 | 0.3700 | 0.1923 | 0.1181 | 0.0925 |
| | $Bias^2$ | 0.0075 | 0.0002 | 0.0002 | 0.0002 | 0.0001 | 0.0001 | 0.0000 |
| PIV | $Var$ | 0.4155 | 0.2869 | 0.1567 | 0.0871 | 0.0429 | 0.0267 | 0.0184 |
| | $MSE$ | 0.4229 | 0.2871 | 0.1570 | 0.0872 | 0.0430 | 0.0268 | 0.0184 |
| $Bias^2_{PIV}/Bias^2_{LS}$ | | 0.0306 | 0.0007 | 0.0013 | 0.0007 | 0.0003 | 0.0003 | 0.0000 |
| $MSE_{PIV}/MSE_{LS}$ | | 1.5988 | 1.0450 | 0.6641 | 0.3304 | 0.1831 | 0.1153 | 0.0908 |
| $Bias^2_{PIV}/Bias^2_{IV}$ | | 0.2185 | 0.0038 | 0.0927 | 0.2667 | 0.3438 | 0.5435 | 0.4745 |
| $MSE_{PIV}/MSE_{IV}$ | | 0.0005 | 0.0138 | 0.4828 | 0.8929 | 0.9520 | 0.9756 | 0.9824 |

Table 2: Integrated squared bias, integrated variance, and integrated MSE of the penalized and unpenalized IV estimators ($\hat{g}_\tau(\cdot)$ and $\hat{g}(\cdot)$).

| $\tau$ | CV Value |
|---|---|
| 0.015 | 37.267 |
| 0.05 | 37.246 |
| 0.1 | 37.286 |
| 0.15 | 37.330 |
| 0.2 | 37.373 |

Table 3: Cross-validation values for the choice of $\tau$.

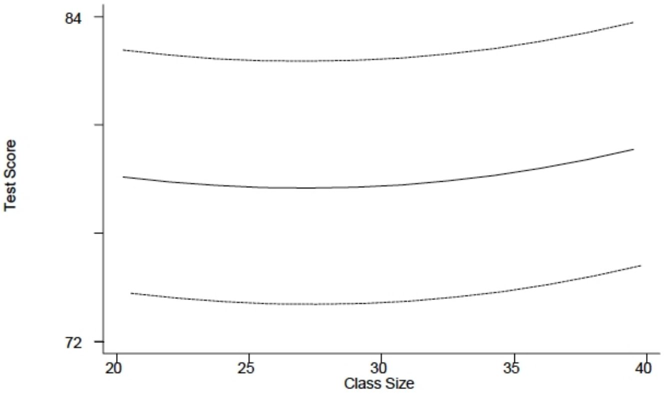

Figure 11: NPIV estimates from Horowitz (2011), full sample ($n = 2019$), 95% confidence band

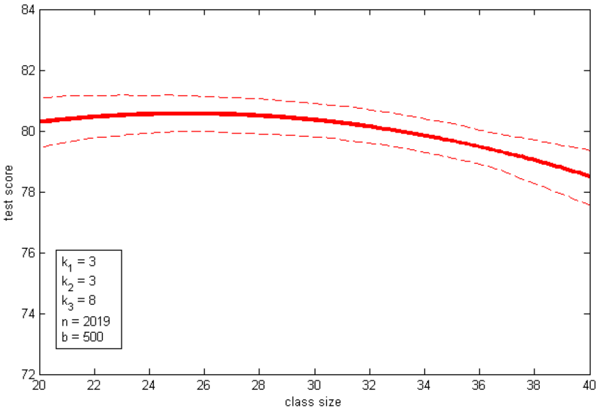

Figure 12: Unpenalized IV estimates with nonparametric first-stage equations, full sample ($n = 2019$), 95% confidence band

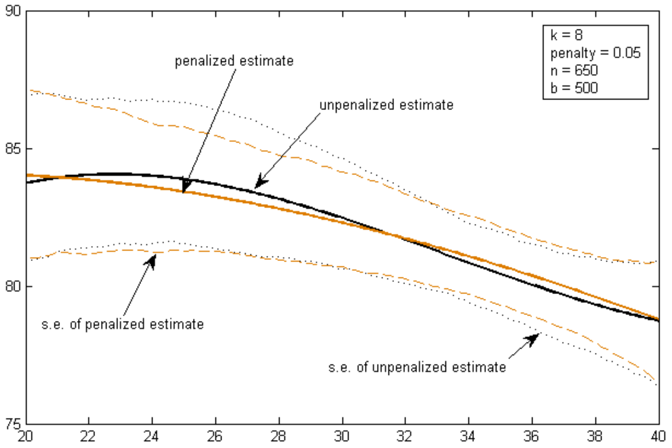

Figure 13: Penalized IV estimates with the discontinuity sample ($F = 191.66$).

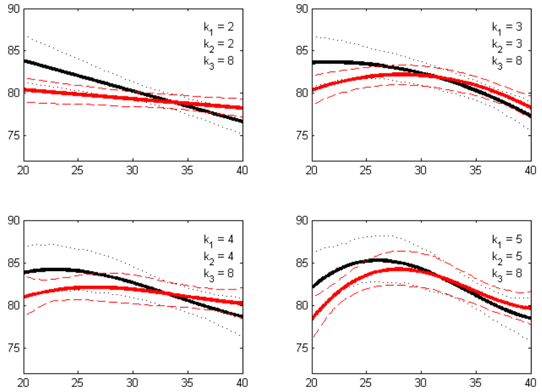

Figure 14: IV estimates with linear vs. nonparametric reduced form ($F = 191.66$).

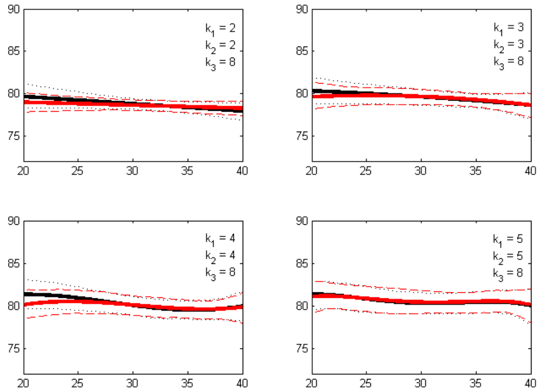

Figure 15: IV estimates with linear vs. nonparametric reduced form ($F \approx 691$).